The Computer Simulation of Syllogism Solving using Restricted Mental Models

Robert Inder

Ph.D.

University of Edinburgh

1987

Acknowledgements

The production of this thesis owes much to many people. In particular I wish to thank Brendan McGonigle for his provocative questions and his continual supply of enthusiasm, not to mention No. 3, and John Lee for his patience in listening to hare-brained ideas. But I must also thank everybody else who may have contributed, including all those who attended the workshops at the School of Epistemics, and of course my supervisors, Henry Thompson and Keith Stenning. Finally, my deepest thanks go to Toog, for waiting so patiently so often while I did just one more paragraph....

I declare that I composed this thesis,
and that the work it reports is my own.

Robert Inder

Table of Contents

Chapter 5: Theories of Syllogistic Reasoning

Chapter 6: Johnson-Laird's Account of Syllogisms

Chapter 7: A New Theory of Syllogistic Reasoning

Chapter 8: Summary and Conclusions

References

Appendix A: Results from the Author's Experiment

Appendix B: Paper Submitted to AISB Conference, 1985

List of Illustrations

Abstract

It is currently believed that the best hope of understanding mental processes, and particularly their predictability in non-physical terms, is to regard them as having much in common with computational systems. Moreover, there is ample evidence that mental life arises from the interaction of a number of independent processing systems, with many of the properties that Fodor associates with modules. This thesis argues for the idea that the communication between these systems should be given a central position in any psychological theory. It also suggests that there are limitations on this communication – specifically, it is restricted to a universally accessible representation of a single situation – an information structure that can be identified with the concept of a Mental Model, as argued for by Johnson-Laird.

A psychological theory should always aim account for observations in terms of the cognitive architecture with the least flexibility. To do so, it is important to impose the tightest possible restrictions on the representational power at its disposal, particularly for communication between its sub-systems. It is suggested that while this is straightforward in the case of world-related phenomena, such as perception and mental imagery, similar restrictions should be be applicable to the information structures used to explain more abstract mental phenomena.

The task that Johnson-Laird uses to illustrate the application of mental models to abstract thought is syllogistic reasoning. The metatheoretical issues that this raises are set out, and the literature on the task is extensively reviewed. In this context, Johnson-Laird's own account of the performance of the task is presented, and criticised for its reliance on extensions to the representational power of mental models.

Finally, a drastically new theory is advanced which uses only the representational power of restricted mental models, and which makes totally different assumptions concerning the nature of the syllogism solving task, which is treated as an exercise in formal problem-solving skills. As such, it is predicted that subjects will each bring different skills and methods to bear on the problem, and attention is focused on capturing the results of individual subjects' reasoning in terms of the interaction between sets of heuristics for generating and defeating conclusions. The results of experiments carried out to support the theory are presented, as are the results of using a computer program to model the performance of individual subjects.

CHAPTER 0

Introduction

0.1. Mind and Body

The term "psychology" is used to cover an enormous diversity of activities. Its problems range from the perceptual abilities of cats to the behavioural problems of children. It probes these with tools that include electrodes and personality tests, and has produced theories in terms of the influence of the social environment, the processing of information in short term memory and statistically significant dispositions to behave. If there is anything common to this spread of activities, it is perhaps best captured by the gross generalisation that psychology is the activity of analysing, with the objective of explaining, human (and indeed animal) behaviour.

While this range of problems and approaches tends to camouflage any underlying unity the field might possess, their very diversity also suggests the outline of one its most fundamental problems. Physicists have made considerable progress in discovering the rules that characterise the behaviour of matter, at least over the range of conditions considered by non-physicists. Within the scientific community, the principles it discovers are recognised as **laws**, at least in the sense, set out by (Fodor, 1975), that they are exceptionless and inviolate. This means that they completely constrain, and thus precisely predict, the behaviour of all matter.[1] Of course, unexpected things do still happen, and the reason for this, and for the existence of other "natural science" disciplines, is that these laws only apply at the level of atoms.[2]

Any observable system contains many orders of magnitude more atoms than could conceivably be individually analysed in terms of the laws of physics. As a result, systems of interest must be analysed in the terms of some other discipline, such as chemistry, thermodynamics, or electronics. These disciplines are nonetheless constrained by the laws of physics, which still apply, and serve only to facilitate their application to interesting situations by identifying appropriate idealisations. Such a process involves neglecting factors which may, under certain circumstances, turn out to be relevant. As a result, these disciplines can offer not exceptionless laws, but only fallible rules, although these might be very accurate and useful.

---

[1] Complications related to necessary uncertainty and quantum effects are, for psychological purposes, eminently ignorable.

[2] Or sub-atomic particles, of one flavour or another. However, the smaller the particles, the larger the energies involved and the fewer the situations where the effects are relevant. For most phenomena, on Earth at least, atoms are atomic.

To see what is involved here, consider the operation of a television set. There is no possibility of even identifying the particles that make up its circuits, let alone applying the fundamental laws of physics to them. However, televisions are manufactured from parts that approximate ideal components – capacitors, transistors etc. – and in terms of these idealisation the performance of the set can be predicted. The cost of doing this is that the mechanical structure of the circuit is ignored, even though eventually, thermal effects and vibration will lead to its deterioration and the failure of the set. Crucially, the way the structure of the circuit leads to failure will be predictable from the laws of physics, and the unpredictability of the breakdown arises from the imperfection of the idealisation.

The existence of exceptionless physical laws for the behaviour of all matter has implications for psychology, since they must govern the bodies of every living creature. Indeed, every individual movement or action can be described in terms of forces and masses, couples and moments, energy and momentum. Doing so reveals motion as the conversion of chemical potential energy into kinetic energy, performed by muscles as predicted by biochemistry. The activity of these muscles is enabled by the electrical activity occurring on certain nerve fibres, which in turn are influenced by others and so on back to the brain. Yet even there those same physical laws, seen here in the guise of neurology, govern every neuron firing. Thus the fundamental laws of physics show how every movement of every creature is caused by the biochemistry and neurology, and thus fundamentally the physics, of its nervous system.

This physicalist's view of bodily action meshes well with observations of how the structure of the sense organs, such as the ears, allows externally specified phenomena, such as air pressure variations, to cause neurons to fire. It naturally predicts the possibility of electrical stimulation of muscles and the correlation of electrical activity in the brain with thought or arousal. It also leads one to expect the kinds of paralysis and loss of sensation that damage to the nervous system produces. More importantly, since the brain is so complex, it offers the possibility of a mechanism that can support the range and flexibility of observed behaviour, although unfortunately this same complexity also means that the details of the operation are currently incomprehensible. Finally, it even shows why the presence of drugs in the brain, or indeed physical damage, can affect behaviour. Thus the recognition that bodies are necessarily governed by the laws of physics leads to a plausible and potentially predictive explanation of behaviour, where the effect of the physical state of the world on the sense organs is seen as triggering the brain to stimulate muscles in a manner determined by its electronic and chemical state.

Nevertheless, the idea is confronted with a very severe problem. The scientist may be happy to see the brain controlling the behaviour of Neddy's body in line with the laws of physics. However, Neddy is likely to be less than happy about it, and protest that it is **he**, Neddy, that controls his body. This is a standard embodiment of the mind-body problem that has amused philosophers since the ancient Greeks. In this form it is confounded by problems relating to consciousness, free will and responsibility, but it can be re-expressed to remove these, whereupon the essential problem becomes clear. Doing so requires only the innocent observation that Neddy seems very likely to protest at a theory that denied that he controlled his body. The point is that

Neddy's annoyance can be predicted very easily using only informal, everyday knowledge of how people behave.

This is what Pylyshyn has in mind when he says

> Although textbook authors are fond of pointing out the errors in folk psychology, it nonetheless far surpasses any current scientific psychology in scope and general accuracy. What is significant about this is the mentalistic vocabulary, or level of description, of folk psychology...
>
> (Pylyshyn, 1980. P112)

Folk psychology predicts a person's behaviour on the basis of his beliefs and desires. Crucially, these predictions are **accurate**. Moreover, this fact is unaffected by whether or not one chooses to say that Neddy actually **has** these beliefs or desires. It is enough that the observed behaviour is regular under this description (which Dennett (1978) suggests is sufficient condition for ascribing them). Thus the core of the dilemma is that the same behaviour must satisfy two different sets of constraints and predictions. It must be compatible with an ascription of beliefs and desires, but as a physical system, it must also, simultaneously, obey the laws of physics to the lowest level.

There are, of course, many cases where we can describe the same events in more than one way. The example above illustrated how electronics terms can be applied to a situation without in any way undermining the applicability of the laws of physics. However, there is a crucial difference between that and the case in question. The categories of electronics – e.g. transistors and capacitors – represent useful idealised approximations of the underlying physical laws and it is clear that, and in what sense, the predictive power arises from them. In this case, electronics can be said to be *reducible* to physics. (See Fodor, 1975. pp 19 - 26)

However, reduction is not always possible. In fact, there is an entire scale of degrees of independence or reducibility between pairs of compatible descriptions. Fodor (1975) provides a full account of the phenomenon, but for illustration it will suffice to consider a simple example of a television set tuned to "The Benny Hill Show". The physicist might talk in terms of the ambient electromagnetic field inducing varying current flows in a pair of conductors, which in turn affect numerous other potentials and currents until eventually an electron beam is deflected inside an evacuated glass tube. Compare this with the kind of description offered by the television engineer in terms of signals and noise, aerials, pictures, frames, lines and blanking. There is obviously a close mapping between the two, and we can readily see how the two views interact. The situation changes when we consider the description given by an ordinary viewer, mentioning Benny Hill, little bald men and cleavages. There is still a discernible mapping back to the engineers' description, possibly via patches of colour, although it would be highly disjunctive and not at all predictive or interesting. Finally, there is the description by the ardent feminist in terms of repetitive humour and degradation of women. While quite possibly a reasonable description of the situation, and in no way contradicting any of the others, it is completely unrelated to that of the physicist.

In the case of predicting behaviour, the way a person considers a situation – possibly in terms of social relationships, objectives, threats etc. – need bear no regular relationship to its physical "reality". Intentional descriptions are **not** reducible to physics. As Fodor (1975, P25) puts it, these terms "cross-classify the physical natural kinds", and thus there can be no principled

mapping between physical laws and psychology. This raises a significant problem, since there are thus two **independent** descriptions of the behaviour of the body and, unlike the case of comedians on television, both claim the power to predict the very same events. One (psychology in terms of beliefs and desires) is de facto accurate. The predictions of the other are currently incalculable, but if physical laws are exceptionless – a belief that underpins the foundations of science – then presumably they too must agree with the observed behaviour. Since the reasons for the two predictions are completely unrelated, the problem is to explain how they can both produce the same predictions – how it is that they both appear to "march in step".

0.2. Computation and Representation

A possible solution to the problem of operating under two distinct descriptions is suggested by observing that the phenomenon is not confined to psychology. The laws of physics, operating in the guise of electronics, allow the operation of a computer to be predicted in terms of electric currents and gate operations. However as Pylyshyn points out, under certain circumstances its behaviour can be predicted under quite different rules: when the computer prints "5" in response to "(PLUS 2 3)", this can be explained by reference to the laws of arithmetic.

> How can the process depend ... both on physical laws and on the abstract properties of numbers? The simple answer is that this happens because both numbers and rules relating numbers are **represented** in the machine as symbolic expressions and programs, and that it is the physical realisation of these representations that determines the machine's behaviour.
>
> (Pylyshyn, 1980. P113)

Using representations in the explanation of behaviour is often thought problematical:

> Nothing is intrinsically a representation of anything; something is a representation only **for** or **to** someone; any representation or system of representations requires at least one **user** of the system who is external to the system. Call such a user an *exempt agent.*
>
> (Dennett, 1978. P101)

Thus if the brain uses representations, there must be at least one exempted agent to interpret them – a *homunculus* – the behaviour of which must also be explained. Some theorists, e.g. Gibson, seem to believe (see Inder, 1986) that doing so must involve further homunculi, and thus leads to an infinite regression. However, careful theorising can avoid this provided the operation of each homunculus can be couched in terms of the interaction of several **simpler** homunculi. Eventually there comes a point where the individual homunculi are so simple that their operation can clearly be ascribed to some kind of mechanistic procedure – they can be *discharged.* Thus the enlightened theorist can employ representations, and be

> Justly unafraid of homunculi, for they are at most just picturesquely described parts of the switching machinery that ensures (sic) the functional roles of the inner messages.
>
> (Dennett, 1978. P102)

It is crucial to notice that the operation of the homunculi at the bottom of such a hierarchy is explained without the need for an exempted agent to interpret representations. While they may manipulate representations, they do not manipulate them **as** representations, but only as physical phenomena of some kind. Only the theorist can recognise them as representations. In particular, it does not matter **how** the symbols encode what they represent, as long as the manipulations of them are formulated in line with the same encoding, and suitably capture the rules of the domain being

represented.

This gives rise to the *computational* view of mind, which has implications that Fodor clarifies:

> I take it that computational processes are both *symbolic* and *formal*. They are symbolic because they are defined over representations, and they are formal because they apply to representations in virtue of (roughly) the *syntax* of the representations.... Formal operations are the ones that are specified without reference to such semantic properties of representations as, for example, truth, reference and meaning.... formal operations apply in terms of the, as it were, shapes of the objects in their domains
>
> (Fodor, 1980. P226 - 227)

Fodor's formality condition arises because the computer (or brain) must (eventually) manipulate representations not as representations, but in line with the laws of physics. As a result its behaviour can only be affected by the physically determinable features of its internal state (and of the activity of its sense organs). Representations can only bring about different effects, and thus different thought processes, in virtue of being physically distinct – of different ''shapes''. Pylyshyn puts this as follows:

> Because a computational process has no access to the actual represented domain itself... it is mandatory, if the rules are to continue to be semantically interpretable (say as rules of arithmetic), that all relevant semantic distinctions be mirrored by syntactic distinctions – i.e. by features intrinsic to the representation itself... Simply put, all and only the syntactically encoded aspects of the represented domain can affect the way a process behaves.
>
> (Pylyshyn, 1980. P113)

The involvement of representations and procedures allows computational psychological accounts to be tied to accounts in terms of beliefs. Representations offer a way of seeing **why** belief psychology can predict the operation of a physical mechanism, and this is absolutely essential:

> It simply will not do as an explanation of, say, why Mary came running out of a smoked-filled (sic) building, to say that there was a certain sequence of expressions computed in her mind according to certain expression-transforming rules. However true that might be, it fails on a number of counts to provide an explanation of Mary's behaviour.
>
> (Pylyshyn, 1980. P161)

Specifically, it fails to show any relation between this behaviour and any other instances of building-evacuation behaviour – such as **walking** out of a smoke **free** building where (alarm) bells are ringing, or indeed danger-avoidance behaviour in general. The computational approach says that belief-talk works because beliefs and goals are descriptions of certain representations and procedures found within the brain.

Viewing cognitive processes as computational highlights an important feature. Pylyshyn explains that

> By separating the semantic and syntactic aspects of cognition, we reduce the problem of accounting for meaningful action to the problem of specifying a mechanism that operates upon meaningless symbol tokens and in doing so carries out the meaningful process being modelled (e.g. arithmetic).... The machine's functioning is completely independent of how the states are interpreted (though it is far from clear whether this would still remain so if it were wired up through transducers to a natural environment).
>
> (Pylyshyn 1980. P114)

Fodor (1980, P230-231) likens such transducers to ''oracles'' writing on the tape of a Turing machine (See Minsky, 1967. Ch. 5 - 7, or Dennett, 1978. Ch 13), and points out that

> As long as we are thinking of mental processes as purely computational, the bearing of environmental information upon such processes is exhausted by the formal character of whatever the oracles write on the tape. In particular, it doesn't matter to such processes whether what the oracles write is **true**; whether, for example, they really are transducers faithfully mirroring the state of the environment, or merely the output end of a typewriter manipulated by a Cartesian demon bent on deceiving the machine. I am saying, in effect, that the formality condition is tantamount to a form of methodological solipsism.
>
> (Fodor, 1980. P231)

The key advantage of the computational view of cognition is that it allows theorising concerning cognitive processes that are **about** something independently of the philosophical thorns that surround the matter of what it is to **be** a representation. The disadvantage, as Pylyshyn (1980) admits people like Fodor (1978) and Dreyfus (1979) are quick to highlight, is that there is a sense in which a computer does not know what it is doing. However, he also points out that all such arguments

> are not simply arguments that machines cannot represent; they are also arguments that **nothing** (including people) can represent. For what **all** such arguments are notably silent on is exactly what it is that people have, and suitably programmed computers necessarily lack, that accounts for the latter's not being entitled to have representations.
>
> (Pylyshyn, 1980. P161)

They serve only to show that nobody understands what it is to represent.

0.3. The Nature of Computation

The discussion so far has focussed on Pylyshyn's observation that computation involves the manipulation of representations in line with the rules of a particular domain. However, as the actual quotations, and particularly the first in the previous section, show, Pylyshyn also suggests that this manipulation happens in line with rules that are themselves represented. Indeed, the underlying assumption of both Fodor and Pylyshyn is that it is this feature that actually distinguishes computations from other processes, which are in some sense just **following the laws of physics**. Fodor makes this explicit when he suggests (1975, P74) that computers, which on a computational view of mind includes the brain, are distinguished by the fact that a "representation of the rules they follow constitutes one of the causal determinants of their behaviour". Thus a VAX is a computer **because** its behaviour is directed by the bit patterns in its memory, which can, in the light of the VAX instruction set, be seen as encoding the rules it is to follow. In contrast, a transistor or television set is just a physical device, which merely behaves the way the laws of physics determine an object of its particular structure and circumstances must.

The consideration of such obviously categorisable examples suggests that this criterion offers a clear characterisation of computational processes. However, this is misleading. Literally everything can be seen as having its behaviour causally affected by a representation of rules, providing its physical structure is recognised as such a representation. This would suggest that a thermostat is nothing less than a computer following the rule that current must flow only between its terminals when the temperature is below a certain limit, a rule represented by the configuration of its parts. Such a suggestion is, of course, quite contrary to the generally accepted use of the term "computer", and any interpretation of a characterisation of computers that supports it is

vacuous.[3]

The problem arises from the failure to restrict what can count as representation, and in the absence of criteria for doing so, requiring it will not provide a principled classification. The uncertainty stems, at least in part, from the focus of the previous discussion. As pointed out above, nothing is intrinsically a representation, and any mention of representing raises, at least implicitly, the matter of an exempt agent. A requirement that rules be represented is meaningless unless it is clear **for whom**. But Fodor's definition says nothing about a process being computational **for** anybody, and Pylyshyn certainly does not see the distinction as a matter of opinion. Both are seeking to characterise an objective distinction, which means that, if the representation of rules is to be relevant, the identity of the exempt agent must be clear.

The obvious candidate – the ideal observer – is problematical. Given a free hand, the physicist's stalwart ambassador will certainly be able to glean the rules that govern the functioning of a device just by examining it. This means that its structure alone must indeed be serving to represent those rules of operation, at least to anyone with sufficient knowledge of physics. But this is no use – thermostats are not computers, so some more restrictive idea of what is a representation of rules is required. Unfortunately the sharp and naturally obvious division between the VAX and television gives a false impression of the difficulty of this task.

To illustrate the problem, consider a range of ways of implementing a specific function,[4] which, being built to the highest philosophical standards, all produce indistinguishable behaviour. Some are unproblematically computational while others are equally obviously not. In every case there are features – voltages or bit patterns – that can be interpreted as representations of the state of the world, which means that their computational status must be distinguished by whether the rules being followed are deemed to be represented.

(a) A dedicated circuit, made of integrated circuit amplifiers (or even transistors), resistors and capacitors soldered onto a printed circuit board. The resulting circuit would be considerably less complex than a television set, and with no apparent claim to be any more a computer, or any less a dumb follower of the laws of physics.

(b) The same circuit, but with "act like a servo" painted on the back.

(c) The same circuit, but with the components physically re-grouped.

(d) The circuit of (c), but with certain circuit board tracks replaced with terminal posts linked by wires.

(e) The circuit of (d), but with the terminal posts replaced by plugs and sockets.

---

[3] No importance attaches to the normal use of "computation" and related terms – it could be that their normal use is incoherent, or covers an uninteresting or disjunctive collection of phenomena. What matters is that Fodor has offered a definition of some property which is exhibited by paradigmatic computational systems. He thinks it is both principled and relevant to surmounting the mind-body problem, and he believes it corresponds to the everyday notion of computation. For now this label will do, and if this later appears to be a mistake (as is suggested below), the importance of the concept is undiminished.

[4] The one in mind is that of a *servo*. This is a means of controlling the behaviour of a system to ensure that it quantitatively follows a command signal. Typically this involves generating a drive signal to the system based not only on the command signal and the current state of the system, but also its history and performance characteristics.

(f) The circuit of (e), with a number of extra groups of components on the circuit board. None of these in any way affect the operation of those in circuit (e). This device is normally known as an analogue computer.[5]

(g) The circuit of (f), but with each wire now including a switch which is closed (i.e. conducting).

(h) The circuit of (g), but with a number of additional switched wires linking arbitrary points, arranged so that all the switches are open. Since these switches and wires are supplied by the local Philosopher's stores, no current flows through any of the wires, and attaching them has no effect on the operation of the circuit.

(i) The circuit if (h), but with each mechanical switch replaced by an electrical one (e.g. a relay) set according to the value of a bit in a register.

(j) A special-purpose device designed for manipulating analogue signals in digital form. What operations are carried out on the signals is determined by bit patterns permanently stored within the chip. Each pattern specifies an arithmetic operation and conditions under which it is to be carried out.

(k) An integrated circuit containing a general purpose micro-processor and RAM which is executing the program in on-chip read-only memory.

(l) The same components as (k), but distributed between several separate integrated circuits.

(m) A BASIC program running on a BBC Micro. Such a system is undeniably computational, with the RAM image of the BASIC program unquestionably constituting an encoding of the rules that is causally directing its behaviour.

A BASIC program on a BBC Micro certainly is computational, and a dedicated circuit is certainly not. This means that one of the transitions between the systems listed above must have been responsible for introducing a representation of the rule to be followed. Whatever criteria are to serve to recognise computations or the "representation of rules" should make it clear which it was.[6]

What appears to be the most qualitative step is the transition between the register-driven analogue computer and the special-purpose digital chip – i.e. (i) -> (j). In both cases the operation of the circuit is causally determined by a pattern of bits, though in the former, the specified processes are applied simultaneously by separate processing units, while in the latter they are carried out sequentially by the same unit. However, it is not clear why this difference should have any bearing on their status as computations. Nor is it obvious how it should be extended to classify the plethora of degrees of parallelism that currently dominate computer science. Finally, even if a clean distinction could be formulated, it would be a disaster for the computational theory of mind. The brain can process large amounts of information very quickly even though its most basic unit, the neuron, is very slow in comparison – Potter (1975) has shown that semantic content can be extracted from images in only 125 mSecs, or only 100 cycles of the fastest neurons.

[5] These are discussed in more detail in section 5 of this chapter.

[6] Of course, these stages could be "out of order", in which case more than one of the modifications would fit this description.

However, neurons are very numerous, which is usually taken to suggest that this speed is achieved by parallelism. Thus if the representation of the rules being followed is defined as requiring any sort of serial processing, current evidence strongly suggests that this will class the brain as non-computational!

Fodor's definition stipulates that a rule representation must be **causally** involved in the production of behaviour. This means that whichever step introduces the relevant representations must alter some feature which affects the operation of the system. This means that writing on the back of the circuit board (b) will obviously make no difference, because it is irrelevant to the behaviour of the circuit. However, so too are all the changes (b) - (i), which means that there is no possibility of separating the dedicated circuit from the digital signal-processing computer. This forces one to attempt to draw a distinction between (j) and (m), all of which would generally be regarded as computational. Something has gone wrong.

There is, however, another possible exempt agent – namely the system itself. Fodor's criterion could be interpreted as distinguishing systems exhibiting behaviour which is causally directed by features that represent the rules to be followed **for the system itself.** The key point here is that it replaces the notion of something being a representation for an ideal observer with that of being a representation for a definite system. Thus the matter comes down to deciding whether anything that causally directs the behaviour of a system is also a representation for it. Since, as has already been mentioned, representation is not a well-understood process, it is not obvious that this is advantageous. However, progress can be made without having to tackle the problem.

Obviously, for a system to treat something as a representation, it must have access to it. Of course, there is a trivial sense in which every system has access to everything that determines its behaviour – that of having its behaviour determined. Borrowing the terminology of computer science, this can be dubbed *execute access*. However, this is not enough – when something is serving to direct the behaviour of a system, it is acting not as a **representation** of the rule being followed, but as the **rule itself.** A system can only treat one of the determinants of its behaviour as a representation if it has some **further** access to it. Thus this interpretation of Fodor's criterion would be met by systems with *write access* to some of the causal determinants of their behaviour.[7]

This interpretation is much closer to the clean, qualitative, distinction that Fodor and Pylyshyn are seeking, although identifying it with computationality is decidedly problematical. On this account a VAX would be a computer because its behaviour is directed by the bit-pattern in its memory **and** it can manipulate that bit-pattern as data. This is (desirably) in clear contrast to a television set or thermostat, which have no such access to any feature of their structure. Nor do analogue computers, which means that by this criterion they are mis-named, although since they are far from paradigmatic cases, omitting them would not cause too much trouble for everyday ideas of computation.

---

[7] Another possibility is systems with only execute and *read* access. However, at least at first glance, there do not seem to be many interesting members of the class of systems which examine the rules of their behaviour but can do nothing to change them.

However, computer programs written in COBOL, Pascal or BASIC are paradigmatically computational. Yet these languages support no access to any kind of procedure specification. Thus they stand discriminated from processes specified in languages such as Prolog or Lisp, which allow programs to both read and alter the procedures that constitute them. In fact, the dividing line is crossed when PEEK and POKE are added to BASIC, since this enables programs to modify themselves. Such a change clearly has no effect on whether or not the procedures described are computational in the everyday sense of the word. A Pascal program undeniably describes a computation, even though it can never have anything that causally affects its behaviour act as any kind of representation for it. The obvious step, then, is to recognise that this interpretation of Fodor's criterion defines a new, potentially interesting, class of mechanisms, henceforth to be dubbed *flexible autonomous systems* or automata.[8]

Since the definition of a flexible autonomous system arose from Fodor's attempt to characterise computation, it seems natural to probe its relevance to the original problem. The circuits (a) to (i) undeniably include instances of both computational and non-computational mechanisms. However, as pointed out above, none of the differences between them actually affects their operation in any way. In fact, all that is happening is that certain features guiding the operation of the different versions are becoming progressively more accessible from outside the system – they are becoming easier for some other agent to modify. Whereas a flexible automaton is able to modify its own behaviour, being a computer is closely related to being able to have externally modifiable behaviour. A computer is a part of a flexible autonomous system which includes its programmer. However, mechanisms are not simply either modifiable or immutable, and indeed an assessment of "modifiability" must be dependent on the abilities and facilities of the modifier – the circuit in (d) is a computer for anyone with a soldering iron. This suggests that "computationality" is likewise a matter of degree, and will not admit to the kind of sharp distinction that Pylyshyn and Fodor seek.

Finally, it is important to evaluate the impact of these ideas upon cognitive science, and in particular on Pylyshyn's bid to defuse the mind-body problem by invoking computationality, considering it to be a sharply defined, objective, quality. However, closer examination has revealed that this is not the case, and it is important to re-consider the situation in the light of this.

Pylyshyn sought to explain how the brain, a physical system, could nonetheless produce behaviour that followed the (physically) arbitrary rules of some other domain. The key suggestion was that the rules could be followed because they were (physically) represented within the system. This gives the system a certain flexibility of behaviour, even within the constraints of the laws of physics. This led Pylyshyn to suggest that human mental processes are computational, in the sense of being directed by represented rules. However, it was shown above that this term is far from the objective, qualitative property that Pylyshyn thought, but is a subjective matter of degree. The real force of his argument comes from the suggestion that the human mind is an autonomous flexible

[8] This term is clearly geared towards systems that have *write access* to some feature which causally direct their behaviour. Those systems mentioned in the previous footnote, which only have *read access*, should perhaps be termed *introspective automata*. A standard microprocessor with its program in ROM could be programmed to be such a system.

system. People are able to follow arbitrary rules because they can write a suitable encoding of them into their brains. This suggests that precise discussion should focus on the "flexible automaton theory of mind", to evoke the idea of a system obeying rules that it has itself created. Nevertheless, the cautious use of the term "computation" still has a useful role, which it will continue to be used to fill, not only for stylistic convenience, but also to maintain and emphasise the ties with familiar ideas.

0.4. The Computational Theory of Mind

It is now commonplace to describe cognitive processes using computational terminology, such as "storage" and "process", and techniques, such as flow diagrams. However, Pylyshyn believes that the real importance of this process has not been recognised. He suggests that "There is no reason why computation ought to be treated as merely a metaphor for cognition, as opposed to a hypothesis about the literal nature of cognition"[9] (Pylyshyn, 1980. P114), and that doing so "carries with it far-reaching consequences". Viewing cognitive phenomena as revealing the operations of a flexible autonomous system adds greatly to the explanatory power of the processes described in psychological theories. Not only do they produce the observed behaviour – what Pylyshyn terms *weak equivalence* – they can be postulated to actually **be** the mechanism that produced it. As a result, such features as the information requirements and complexity of the task can be determined by direct examination of the process. Similarly, computer programs can be suggested to behave in the same way as subjects because they instantiate the very same computations – exhibiting what Pylyshyn calls *strong equivalence*.

However, there is a price to be paid for this greater power.

> The theory-builder ... is no longer free to appeal to the existence of unspecified similarities between his theoretical account and the phenomena he is addressing.
>
> (Pylyshyn, 1980. P115)

In particular the theorist is committed to postulating a consistent (i.e. specifiable, though possibly unspecified) mapping between any representations he employs and some kind of physical instantiation within the brain. Moreover, this applies not only to the "input" and the "output" of the process, but also to any **intermediate** representations and the transitions between them. Any particular mapping from input to output defines an infinite class of algorithms. The significance that a literally computational view of mind places upon intermediate states and steps means that it is sensible, and important, to ask which one is actually **right**.

There is another facet to this issue. While it is possible to specify a process in a totally abstract manner, carrying it out requires a mechanism. In particular, computations, which of course includes the operations of flexible autonomous systems, need a computer. They can only happen in an environment which has the physical properties that allow the application of suitably formulated transformations and tests to encoded representations. A digital computer is one such mechanism –

[9] Pylyshyn suggests that recognising how a theoretical concept applies to real problems has been responsible for many conceptual advances. He cites the way the application of Euclidean geometry to the world was responsible for the idea of space as "that empty ... three-dimensional receptacle, whose existence and properties are quite independent of the earth or any other object. Such a strange idea was literally unthinkable before the seventeenth century".

its hardware is so arranged as to allow representations, in terms of bit patterns, to be transformed by procedures specified in terms of instructions. If cognitive processes are computations, then the brain, too, must be so arranged that the arbitrary regularities or rules of the domain being considered (i.e. thought about) can shape its operation – it must implement a suitable architecture. Moreover, the architecture available can constrain the range of algorithms that can be executed and the effort they require. Pylyshyn cites the example of searching using *binary chopping* (see Knuth, 1973. Ch. 6), which can only be sensibly employed on an architecture that supports explicit arithmetic manipulation of **addresses** of data. This means that the theorist should worry not only about the algorithms of cognition, but also what engine the brain provides to execute them. The central thrust of Pylyshyn (1980) is to suggest that it is essential to distinguish the algorithms of cognition from the architecture that supports them.

> Both the underlying architecture of the computational model, and the properties of the algorithm or task that enables the system to successfully carry it out, must be explicitly addressed in the explanatory story, and both must be independently justified.
>
> (Pylyshyn, 1980. P122)

Unfortunately, the brain is far too complex for its computational architecture to be determined by suitably applying scalpel and electrode. Indeed, ignorance in this area is so profound that it is not even certain what kinds of physical features are involved. It thus seems unlikely that physiologists will be able to contribute to the identification of the functional architecture for a considerable time.

However, this is not the end of the matter. Since "mental architecture can be viewed as just those functions or basic operations of mental processing that are ... functions instantiated in the biological medium (Pylyshyn, 1980. P126)" then as Pylyshyn points out, the architecture can be defined in terms of these operations – the very features that make its characterisation important. Even though the architecture arises from the biology of the brain, it can still be described computationally or even in terms of homunculi, without necessarily involving a vicious regress (see above). Thus Pylyshyn suggests that psychological theories should postulate both the algorithm that underlies the behaviour and the functions performed by the architecture on which it runs.

The most widely mentioned argument against this approach is Anderson's *mimicry theorem.* This rests on the notion that any observed pattern of behaviour could be produced by any number of different representational systems and corresponding processes for manipulating them. Anderson (1978) offers a proof of this theorem by proposing a procedure that guarantees at least one way of producing the behaviour predicted of one system from any other that is sufficiently flexible. All that is necessary is to arrange that the mimicing system carries out, or copies, the precise steps of the system it is to mimic. Thus, the argument runs, since any observed behaviour could result from arbitrarily different architectures, no weight can be attached to the objective of identifying the true cognitive architecture.

This kind of mimicry is both commonplace and important in computer science, where it is known as *emulation.* It is the principle that gives the Universal Turing Machine (Minsky, 1967. Ch 7) its power, and is even exhibited by Pylyshyn's own example of Lisp as a functional architecture, which seldom has any reflection whatsoever in the hardware of the actual machines that support it.

Instead, standard computers are used to emulate the functional architecture of the language, and Lisp programs are executed in the resulting *virtual* architecture.

There is no reason to doubt any suggestion that emulation might be occur< within the brain – indeed, the fact that a programmer can "desk check" a computer program serves to demonstrate that it can. However, Pylyshyn argues that this is not enough to undermine attempts to determine the cognitive architecture. While emulation undeniably allows any architecture to mimic the input-output mapping of any other – to be **weakly** equivalent to it – it does not make the two computations indistinguishable, or even **strongly** equivalent. An emulator achieves its mimicry by undertaking additional computation, and this is often manifest in the performance of the emulation. A von Neumann based machine can emulate the architecture proposed by Fahlman (1979) and mimic its set intersection results. However, while Fahlman's machine is unaffected by set sizes, a von Neumann emulation is crippled by large sets, and the difference in underlying architectures soon becomes apparent.

This means that, Anderson's arguments about emulation notwithstanding, proposed cognitive architectures that are weakly equivalent can be distinguished because they can be compared on criteria other than the qualitative behaviour they produce. In particular, even though different systems may produce the same behaviour, they will require different resources (e.g. memory) and take different times to do so. Thus although emulation can easily produce weakly equivalent processes, they will typically exhibit different temporal profiles and patterns of failure due to resource limitations. Hence, Pylyshyn points out, the study of reaction times and error distribution patterns offers a tool for distinguishing proposed cognitive mechanisms that Anderson's argument suggests would be indistinguishable.

Nonetheless, even though these features constrain the range of possible cognitive architectures, they are still not adequate to determine the computational system that produced them. Failing to recognise this is a serious and widespread methodological error, since it can lead theorists to attempt to explain a psychological phenomenon by simply constructing an algorithm and architecture that produce the observed data and response times. However, Pylyshyn points out that there are other factors that are affected by the relevant choices, and the necessity of taking account of them is what he has in minds when he is at pains to insist that the design of cognitive models must be *principled.*

Pylyshyn suggests a number of factors that are relevant to identifying both algorithm and architecture. The most obvious of these is the possibility of obtaining evidence relating to the intermediate states through which computation passes. However, many of the others are motivated by his assumption of a clear distinction between computational and non-computational processes. Since the functions the cognitive architecture provides underpin and explain the computational nature of the mind, they cannot themselves be computational, on pain of regress. Hence, it must arise from the fundamental structure of the brain, which, being biologically determined, must be (comparatively) both universal and constant. The features it presents will remain the same when considering different trials, tasks and individuals, each of which offers a powerful methodological constraint. This means that, just as the architecture of a computer is unaffected by the program

being run, so too will the basic functions of the cognitive architecture be immutable by the demands of the task and the beliefs of the subject[10] – they will not be what Pylyshyn calls *cognitively penetrable*. Moreover, different experiments should be explained in terms of revealing features of the same cognitive architecture, while results from different individuals can at least be expected to be similar.

Pylyshyn also attempts to go beyond the obvious implications of this. He suggests the notion that cognitive phenomena are computational "entails the claim that functions attributable to the functional architecture can consume only constant resources" (Pylyshyn, 1980. P127). However, this needs a more explicit justification, particularly since he is including processing time as a resource, as is revealed by his reference to his argument against Anderson's mimicry theorem. Anderson (1978, P266) suggests that the extra computation involved in emulation would be undetectable if the emulating system could operate fast enough. In his response, Pylyshyn (1979, P391) is less concerned with the overall time taken than with the temporal **profile**. The relative difficulty and duration of various tasks is very highly characteristic of a computational architecture, and will not be shared by an emulator based on a different architecture. Pylyshyn wants to reject any attempt to mask the effects of this by varying the speed of the emulating system. Clearly, any pattern of relative timings can be readily achieved if the emulator is able to vary its speed of operation – to "hurry" on the operations that require a lot of computation to emulate or to "dawdle" on the easy ones. However, Pylyshyn emphasises that duration (and resource requirements) are not just additional "outputs" from computational processes, but are fundamental properties of its execution. Thus he sought to establish a firm criterion that would allow the principled rejection of proposals involving such mechanisms. This lead him to suggest that each function of the cognitive architecture always takes constant time. Presumably similar arguments would go through for the use of resources.

In making this requirement, Pylyshyn is imposing tighter constraints than are necessary to achieve his purpose. A weaker criterion can serve as the basis for principled objections to theories that achieve resource or time profiles by modifying an emulation. All that is necessary is that the postulated relationship between a function's time and resource requirements and its operands be independently justifiable. The simplest such relationship, and thus the one that gives the simplest theories, is when these requirements are simply constant. However, seeking this degree of simplicity is merely a methodological desideratum, which Pylyshyn is elevating to the status of a meta-theoretical requirement. In doing so, he is imposing a tight constraint on the functions that can acceptably be ascribed to the cognitive architecture.

Not least among the effects of this is the rejection of any kind of emulated architecture, despite the benefits that the concept of a virtual machine has brought to computer science. There, the objective is to allow the clear expression of a process, and thereby to facilitate its comprehension and study. For such purposes, the validity of a functional architecture is unaffected

[10] The situation with machines with programmable microcode is more complex. Nonetheless there is still a layer of architecture that cannot be modified without a soldering iron, although it is employed solely to provide another architecture, an example of emulation.

by how, or indeed whether, it is physically instantiated. However, Pylyshyn believes that cognitive science must do more than analyse the algorithms of cognition. It must seek those functions provided by the structure of the brain itself that allow it to support the computational nature of mind. Thus he deliberately aims to establish criteria that will reject any kind of emulated architecture, and thereby force cognitive science to probe deeper, towards the non-computational operations of the brain that enable it to constitute a computational architecture. The requirement that primitive operations consume constant resources is intended to be just such a criterion.

However, this constraint appears to be too strong for Pylyshyn's purpose, particularly since it rejects not only any emulated architecture, but also many of the existing (or easily imaginable) artifacts of computer science. For instance, most storage formats for floating point numbers require that operands be aligned (i.e. have the same exponent) before they can be added. If a CPU contains an arithmetic unit that performs the necessary shifting one position at a time, the time it takes to perform an addition will depend on which numbers were involved. Similarly, some common micro-processors have block move and memory search instructions, the duration of which is obviously dependent on the size of block being moved and whether or not the search is successful (i.e. what is being sought). Another example is the time taken to read a block from disk, which depends on the amount of head movement required, and thus on which block is being read. Yet disk operation is surely a function of the architecture.[11] Once again, the duration of an operation of the architecture is affected by the data on which it works – which block is being read. When discussing the operation of familiar computing engines, such variable-duration operations would unquestionably be ascribed to the operation of the architecture. There would be no suggestion that operations such as directing disk head movements should be thought of as part of any algorithm being executed. Thus Pylyshyn's attempt to focus upon the non-computational operations that underpin the mind have led him to impose a constraint that is significantly stronger than those that dominate computer science.

Moreover, even Pylyshyn's objective itself is of questionable merit. Virtual architectures are indispensable for describing behaviour within computing systems. To argue that the description of cognitive phenomena should avoid admitting any kind of emulated cognitive architecture is equivalent to suggesting that programs running in a Prolog interpreter can best be understood at the level of register transfers within the CPU. This is in stark contrast to the way the practitioners of computer science behave. Even though they are perfectly aware of the truth of the situation, they persist in describing Prolog programs as running within a functional architecture defined by the language itself. What allows them to do this is that the performance of the system is minimally affected by the physically instantiated functional architecture. What makes them do it is that to do otherwise would be to involve so much complication as to completely obscure the behaviour of interest. Within the computational architecture defined by Prolog, reading a term is a primitive

[11] In fact in most computers the processor continues running another task while the disk transfer is achieved by a disk controller which signals when it has finished. However, it is not hard to imagine a computer in which the processor was held in a wait state until the task was complete. Combining such a machine with hardware virtual memory would result in a processor where the execution time of every instruction could depend on its operand. Indeed the effect would appear with any form of non-homogeneous memory.

operation. To abandon this viewpoint would be to render even the simplest program literally beyond human comprehension, obscured amongst the reality of unification algorithms, memory allocation, parsing, ascii codes, buffers and pointers, block sizes, disk head movements, parity checking, time sharing and paging. There is no reason to believe that these factors will not apply every bit as forcefully to attempts to understand the operation of the mind within the brain, or that to proscribe the conceptual crutch of the emulated architecture will have any less crippling effects on attempts at comprehension.

Finally, and most fundamentally, it must be recognised that, as pointed out above, rule-following is not the clear-cut property that Pylyshyn assumes. It was suggested instead that the key notion is that cognitive processes constitute an autonomous flexible system, and it is necessary to re-examine his ideas in this light. Once "computationality" is recognised as a matter of relative degree, the requirement that the cognitive architecture not be itself computational loses its meaning. This is not to say that Pylyshyn has not highlighted an important property. Any system that might be called computational, including an autonomous flexible system, manipulates representations – data – in line with some kind of represented rule – program. Moreover, every such system requires a mechanism with the appropriate kind of flexibility to be able to have its behaviour directed by these rules – a computer. There must still be a cognitive architecture. However, there is no sense in which it can be declared to be non-computational, nor any way in which "emulated" architectures can be identified or rejected.

Nevertheless, the cognitive architecture must still exhibit many of the properties that Pylyshyn required of it. Even when its computational status can no longer be debated, it must still be universal, at least to the extent that cognitive psychology is worthwhile. Less obviously, perhaps, it must still be cognitively impenetrable – unaffected by cognitive phenomena. Even though the flexibility of the mechanisms that create the functional architecture cannot be restricted by Pylyshyn's ideas of computationality, the features that direct its implementation must be inaccessible to the processes it supports. Were they not, the proposed cognitive architecture would just be part of the autonomous flexible system that it was required to underpin, the architecture of which would, of course, still require explanation.

Furthermore, when considered over a longer period, a similar argument shows that the cognitive architecture must still be (relatively) constant between tasks and over time. Of course it is by no means necessary – indeed, it is highly unlikely – that the cognitive architecture of an individual is absolutely constant. At the very least growth and maturation are highly plausible sources of change that were present even within architectures that met Pylyshyn's tight constraints on acceptability.[12] But losing those constraints increases the range of possible implementations of the cognitive architecture, and thus the ways that other mechanisms – such as learning – can alter it. Cognitive impenetrability only rules out modification by cognitive processes whereas, for example, certain features of the cognitive architecture might even be implemented by independent,

[12] Since the fabric of the cognitive architecture – the brain – must grow, even Pylyshyn would not suggest that the functions it implements will not change.

self-contained, flexible autonomous systems.[13] However, to the extent that these changes are gradual or infrequent, cognitive psychology, of adults at least, can simply ignore them. Whereas to the extent that they are observable, they will have to be treated as cognitive phenomena, and the cognitive architecture extended to embrace both the systems that provoke them and those that make them possible.

Pylyshyn sought to advance criteria for the identification of the one true cognitive architecture that made it possible for the brain to support computational processes. However, he could only do this on the assumption that the world could be clearly categorised into "computational" and "non-computational" processes. Recognising that it cannot means that cognitive science must be prepared deal explicitly with the fact that cognition may well be supported by layers of emulated architectures. However, to admit this is merely to confront the true complexity of the task. Crucially, the possible involvement of emulations forces no new kinds of indeterminacy into cognitive theories. They will indeed be underdetermined by the available evidence, but no more than those of any other science, so that references to such underdetermination are best dismissed as "true, but uninteresting". There is nothing that should change Pylyshyn's opinion of the matter:

> It is not clear to me why people seem to feel that inferring the nature of a cognitive process is any more subject to inductive indeterminacy than is, say, inferring the nature of a chemical process, except perhaps that cognitive phenomena may themselves not be well enough specified at present.
>
> (Pylyshyn, 1980. P163)

By likening cognitive and chemical theories, Pylyshyn stresses that cognitive theories must be subject to the same evaluations as those in other sciences. A postulated cognitive architecture should be regarded in the same light as a hypothesised chemical structure – every bit as "real" and every bit as testable. Indeed Chomsky (1980, P189ff) presents a similar argument, likening the process of characterising the mechanisms of human language to to the determination of the physics at the centre of the Sun. Further, as with other sciences, this testing is not just in terms of empirical accuracy – ability to match the data – but also involves metatheoretical criteria such as elegance and parsimony, range and predictive power, and consistency with other theories.[14]

---

[13] Paging systems already exist that decide which pages to keep in main memory by applying non-trivial algorithms to patterns of usage. The relevant information is gathered and manipulated by the hardware, and, as with the rest of the paging process, is completely invisible to the computational processes being supported.

[14] Indeed, the factors that Pylyshyn mentions – reaction times, error patterns and capacity limitations – are not the only ones that can be brought to bear when evaluating the plausibility of a proposed cognitive architecture. In particular it is significant that the structures that implement this architecture were not designed, but arose as a result of evolution. This is not to suggest that there is a limit to their complexity – the intricacy of many of the mechanisms essential to life is breathtaking. However, it does mean that the mechanism must have arisen incrementally – there was no one monumental chromosomal rearrangement suddenly created the blue-print to grow a working Turing machine in every skull. Furthermore, each step must have produced a working system that conferred survival advantages and would have taken no explicit account of extensibility – expansion busses only appear because a designer decides their cost will be repaid by future developments. Similarly the system could only have developed algorithms for which information was available to both shape and evaluate the output.

Moreover, there is every reason to expect much of this history to be apparent. Once a working architecture appeared, there would be no survival benefit in (and hence evolutionary pressure towards) duplicating its performance. A new system, however superior it might be in terms of efficiency, size or expandability, would bring no advantage – indeed, it would be a liability – until it developed to the point where it has achieved as much functionality and reliability as the old. Experience in applied computing shows that, in the short term, a given amount of development will produce more gains if applied to extending or patching existing software. Effort is only invested in reimplementation because the designers experience suggests that it will be repaid in the long term. This is not the kind of factor that influences evolution – bipedal human beings have the same bones as bears and bats, and that is why they have lumbago, not to mention appendicitis.

When considering the cognitive architecture, it is important not to be too influenced by experience of existing computer science. The physical architecture of almost all existing machines are specific implementations of the same conceptual machine – that proposed in 1946 by von Neumann. They all support a range of operations that rely on the ability to retrieve information from an address. However, as Pylyshyn points out

> This whole array of formal properties is available in all common computer architectures, because they all use numerical expressions for register (i.e. place) names and have built-in primitive arithmetic operations. But these are part of such architectures for reasons that have nothing to do with the needs of cognitive science.
>
> (Pylyshyn, 1980. P126)

Similarly the virtual architectures offered by most programming languages have been designed to facilitate efficient mathematics on these architectures.

> It is important for the use of computational models in an explanatory mode... that we not take certain architectural features for granted simply because they happen to be available in our computer language.
>
> (ibid)

One must be very wary of concepts such as addresses, pointers, loops and counters. These really only have meaning in a von Neumann architecture, when radically different architectures are possible which support entirely different primitive operations. Pylyshyn cites the example of that proposed by Fahlman(1979), which is able to carry out set intersections in a fixed number of steps using *spreading activation*, a notion originally proposed by Collins and Quillian, but better described by Collins and Loftus (1975). Similarly, analogue computers are a long established family of architectures that exhibit many features quite unlike the familiar digital computer, and what little is known about how the brain manipulates information suggests that these may be very relevant to understanding it.

An electronic[15] analogue computer consists of a large number of independent processing units – adders, multipliers, integrators etc. Each of these is sensitive to the electrical activity at its inputs (usually the voltage applied to them), and continuously forces the electrical state of its output (again usually the voltage) to reflect some defined function of the inputs. Calculations are "programmed" by interconnecting the inputs and outputs of combinations of these units. Thus to calculate, for instance, two quantities given their sum and their difference (as for decoding a stereo program in standard FM radio transmissions) one would connect both signals (a + b and a - b) to the inputs of an adder (to get 2 X a) and to the inputs of a subtractor (to get 2 X b). Each of these would be connected to separate divide-by-two units, the outputs of which would represent the decoded signal.

There are three important points to notice about such an arrangement. The first is that the usual distinctions between processor, hardware, program and data break down. There is no distinction between the procedure for doing something – the knowledge of how to do it – and the machine that actually does it. Certainly, as was mentioned above, there is no sign of any explicit

[15] Mechanical analogue computers are also possible (and historically significant). Currents and voltages are replaced by rotations and positions of interlocking gears and wheels – see Hollingdale and Tootill (1965, Ch. 4 & 5). The following remarks also apply to such machines.

representation of the rules that guide behaviour which Fodor suggests define computation. The second is that once a program is written , which involves the assignment of hardware to the task, there is no need to allocate further resources to "run" it. Finally, the delay in actually carrying out a computation is negligible – the decoded version is available virtually as soon as the coded version is presented. Moreover, the output is always "up-to-date", in that it continually changes to reflect the decoding of the current state of the input. Together these mean that once an analogue computer is able to do something – hardware has been allocated to a task – actually doing it is effectively instantaneous and effortless. In short, an analogue computer can do everything it knows how to do, simultaneously and all the time. This is, of course, in sharp contrast with many of the intuitions fostered by experience with sequential machines, which suggest that doing things is expensive in terms of time and resources.

Cognitive science must accept the possibility that it is studying a flexible autonomous system that is supported by an architecture far removed from anything that computer science is familiar with. This may suggest that the concepts that form the basis of that discipline – program, data, storage, flow of control – may be quite inappropriate to the task in hand. Nonetheless, there is an overwhelming reason why the kind of flexible behaviour that Fodor and Pylyshyn tried to label as "computation" should be seen as central to the explanation of cognition.

> The most fundamental reason why cognition ought to be viewed as computation... rests on the fact that computation is the only worked-out view of process that is both compatible with a materialist view of how a process is realised and that attributes the behaviour of the process to the operation of rules upon representations.
>
> (Pylyshyn, 1980. P113)

In short, computation is currently the only conceptual tool available that looks appropriate to the task of describing the gross features of the dynamic behaviour of very complex systems, a situation summed up by Pylyshyn's endorsement (P159) of Fodor's view that it is "the only straw afloat".

0.5. The Aims of Cognitive Science

At this point it is worth reviewing the style of psychology which Pylyshyn has outlined, and which he believes constitute the foundations of cognitive science. The structure of the brain implements a computational architecture, geared towards the flexible manipulation of representations. The physical mechanisms underlying these representations is not known, but this does not matter because it does not actually affect the computations. Mental processes can be identified with computations executed within the architecture, and indeed

> The analysis we have been giving can be viewed as setting out a particular program of research – namely that of designing a cognitive virtual machine, or rather a system or programming language having a functional architecture appropriate for implementing cognitive algorithms... every algorithm that could be executed on this virtual machine would now be considered a humanly possible cognitive process... All and only cognitively possible (though for various reasons perhaps not actual) algorithms and representations are permitted.
>
> (Pylyshyn, 1980. P128 - 129)

The cognitive architecture of the mind plays a similar role in this view of Cognitive Science as Universal Grammar plays in linguistics (See Chomsky, 1980 P28 - 30), and its determination should be a correspondingly central task.

As pointed out above, many factors are relevant to determining the cognitive architecture. Nonetheless, it remains underdetermined, and Pylyshyn (1980) argues for the metatheoretical principle that theorists should postulate the weakest possible architectures, even at the cost of a computationally awkward system, because "the more constrained a notation or architecture, the greater the explanatory power of resulting models" (P126). Furthermore, the algorithms produced by this strategy "Committed one to the weakest presuppositions about the underlying architecture" (P129), and are thus most tolerant of its revision. Since the functional architecture is the point where the theorising of cognitive science meets the explorations of biology, flexibility here will surely ease the job of accommodation. Similarly, simplicity will surely make the evolution of the architecture easier to explain.

Determining the power of a proposed cognitive architecture is not trivial. It is not simply a matter of evaluating the complexity of the primitive operations it supports, since complexities could only be assessed with respect to a means of implementation, the specification of which is currently beyond the scope of brain science. The most obvious ways are by restricting the kinds of things that it can represent and the range of primitive operations that it can invoke on them.

Since such restrictions are desirable, it is of particular significance that it is possible to justify the introduction of a number of them by considering work motivated by phenomena involving mental images. Pylyshyn himself considers that "Imagery is a pervasive form of experience and is clearly of utmost importance to humans" (Pylyshyn, 1973. P2). This leads him to recognise that a natural way of weakening the cognitive architecture would be to assume that these images directly reveal its most basic functioning. He suggests that scenes in the world are represented and manipulated in an *analogue* form, commenting that

> Analogue representations, as generally conceived, are not articulated symbolic expressions, and analogue processes are not viewed as rule-governed symbolic computations. They can thus be viewed as characteristics of the functional architecture. I believe that it is this quality of incorporating fixed constraints into the architecture, and therefore weakening the expressive power and consequently increasing the explanatory value of the models, that makes analogue systems particularly attractive in cognitive science.
>
> (Pylyshyn, 1980. P130)

Theories that do this are currently attracting a lot of attention within cognitive science, though to Pylyshyn's chagrin their proponents seldom justify their enthusiasm so explicitly.

To fully appreciate Pylyshyn's position, it is necessary to recognise the importance that he attaches to the fact that "analogue processes are not viewed as rule-governed symbolic computations". Once again he is invoking the sharp characterisation of computational processes, this time to distinguish theories of mind based on propositional representations of the world from those that involve imagistic structures. The former are manipulated computationally, in line with explicitly represented rules, while the latter are merely shaped by the laws of physics acting upon the structures of the brain. As such, analogue representations are qualitatively less flexible, not least because the mechanisms for dealing with them are biologically determined, and thus immutable. As a result, theories that postulate them make stronger predictions, because they sacrifice all flexibility in their manipulation. In particular, requiring that representations of the world only be manipulated by the fixed, non-computational processes of the cognitive architecture

would severely restrict the possible operations.

Despite having admitted the methodological desirability of constraining the cognitive architecture by ascribing a causal role to mental imagery, Pylyshyn rejects the possibility of doing so. His principle reason is that he believes there is evidence that the complex operations involved are cognitively penetrable, which of course excludes them from the functional architecture. However, his arguments are unconvincing – an issue that will be developed in the next chapter – because his notion of cognitive penetrability rests on his questionable assumptions concerning the cognitive architecture. In particular, his attempt to distinguish propositional representations by the nature of the processes that manipulate them founders because the computational status of the processes cannot be cleanly determined.

The conclusion that there is no principled way of deciding the nature of a representation was, in fact, the main result of Anderson's (1978) indeterminacy arguments. He maintained that the representation of information could **only** be discussed in the context of the procedures that manipulated it. Thus he believed it completely meaningless to debate the format in which information is stored, and in particular whether it is encoded propositionally or imagistically. He supported this suggestion by showing that the behaviour of any process manipulating imagistic representations could be mimiced by another based on a suitable propositional encoding. As described above, he simply argued that the mimicing process need merely emulate the original, which trivially guarantees that it brings about the right behaviour. The only condition is that the propositional representation must be *distinction preserving* with respect to the image. It must capture every feature that the image system can distinguish, which means it doesn't lose any useful information that is in it. Thus although the **content** of mental representations could be determined, Anderson argued that there is no sense in which its **format** can be sensibly discussed.

As described above, Pylyshyn argued that the behaviour it normally produces is far from the only thing that can shed light on the operation of an information processing system. In particular, the time and resources it requires, and the way it malfunctions can both reveal a great deal about the relative accessibility of stored and inferred information. However, this still does not warrant attempts to distinguish propositional and imagistic representations. The point is not that they cannot be **empirically** distinguished, but that they cannot be **conceptually** distinguished, at least by any kind of appeal to the computational nature of the processes that manipulate them. Thus although Pylyshyn has refuted Anderson's supporting argument, his conclusion concerning the futility of debating the imagistic nature of mental representations holds.

In the absence of a sound reasons for rejecting them, Pylyshyn's observation of the methodological desirability of "analogue", or imagistic, theories demands that they be given further attention. The next chapter concentrates on the evidence that leads people to interpret imagery as the result of an analogue process or as some other direct manifestation of the lowest levels of the functional architecture. It concludes, like Pylyshyn, that the kind of two-dimensional array of "surface" information is inadequate to account for many observations. However, it also proposes that a richer representation of the state of the world, such as that provided by the input modules of Fodor (1983), can contribute to a cognitive architecture that is able to capture the data

and the intuitions of the work on imagery. Chapter 2 then elaborates the kind of architecture that can make explanatory use of these enriched information structures, which are identified with Johnson-Laird's notion of a *mental model*. In particular, it stresses not only the need to see mental life as arising from many simultaneously operating information processing systems, but also the vital importance of the communication between those systems. In particular, it suggests that representations of the kind that underpin mental imagery constitute the primary medium for inter-communication between autonomous processing systems.

While it is comparatively easy to see the applicability of such representations to perceptual (or in the case of imagery, perhaps pseudo-perceptual) processes, any proposed cognitive architecture will have implications for more abstract cognitive processes. Chapter 3 begins to broach this topic by extending Johnson-Laird's arguments for the use of mental models in language perception, while the rest of this thesis presents, rejects and reconsiders his account of their use in solving categorial syllogisms, a particular form of abstract reasoning. In particular, Chapter 4 considers the meta-theoretical issues pertaining to a theory of syllogistic reasoning, and Chapter 5 reviews the relevant experimental and theoretical psychology prior to Johnson-Laird's own work. Chapter 6 then presents Johnson-Laird's own experiments and theories, which are criticised, primarily because they rely on an otherwise unjustified extension of the representational power of the mental model. Chapter 7 then proposes an alternative and superior account that paints a completely different picture of the fundamental causes of the unsound reasoning exhibited by untrained subjects. Finally, Chapter 8 summarises the main features of the proposed theory of mental models.

CHAPTER 1

Mental Imagery

1.1. Imagistic Theories

According to the widely accepted arguments outlined in the previous chapter, the kind of flexibility that underlies the power and generality of the computer plays an essential role in cognitive phenomena. As a result, cognitive science is presented with two complementary objectives – the description of cognitive behaviour in terms of the execution of algorithms, and the characterisation of the functional architecture that the brain provides to support this process. Sound methodology for pursuing these ends involves seeking the weakest, most constrained, architecture possible. In this context, Pylyshyn (1980) points out the potential merit of certain theories that ascribe a causal role to mental imagery, because in doing so they suggest many severe constraints on the cognitive architecture. The existence of imagery phenomena is now undisputed – it is universally accepted that people do experience sensations akin to those resulting from perception, but which are unconnected with any interaction with the world. However, the issue of the importance that can be attached to this fact remains highly controversial. A considerable body of literature has been directed towards its resolution,[1] with so little resulting consensus it has been suggested that

> One of the most striking aspects of the current debate over the status of mental imagery is that so much has been written with so little effect.
>
> (Keenen and Olson, commentary on Kosslyn et al., 1979)

There is now substantial experimental evidence which suggests that mental imagery shares many of the same capacity limitations as visual perception. Kosslyn (1975) reports that people find subjectively smaller images harder to handle, and that they need to approach or "zoom in" on them reveal their details. Moreover, he suggests (Kosslyn et al., 1979) that this process causes images to "overflow" round the "edges" at a constant visual angle. Similarly, Finke (1986) describes experiments that suggest that images loose resolution with increasing angle from the point of focus in the same regular manner as perception. He also describes experiments which show that suitable imaging beforehand can enhance performance in perception experiments. These results together suggest that mental imagery both shares subjective similarities and has objectively observable interactions with perceptual processes. These phenomena are most often explained by

[1] Indeed, there is even a journal devoted to the topic: *The Journal of Mental Imagery*. Most of the references cited contain substantial reference sections, particularly Pinker (1984). Finke (1986) and Cooper and Shepard (1984) present a very accessible outline of a great deal of the pro-imagery evidence.

suggesting that it involves part of the perceptual apparatus.

However, in addition to this connection with the mechanisms of perception, other experimental results reveal similarities with the processes that are being imagined. Possibly the best known of these are those of Shepard et al. – e.g. Shepard and Metzler, 1971, Cooper and Shepard, 1973. These experiments require subjects to determine whether or not two images presented to them are views of a single shape or object in different orientations. The results show that the time that subjects take to recognise those cases where they are is proportional to the angle between the two orientations. The proposed explanation of this phenomenon is that subjects tackle the task by rotating a mental copy of the image presented to them (or a model of the object it represents), and the observed effect arises from the fact that this rotation is carried out at a constant rate. Similarly, Kosslyn Ball and Reiser (1978) and Finke and Pinker (1983) offer experimental evidence that the time subjects take to perform a task involving two points within a mental image is proportional to the distance between those points. Here, obviously, the suggestion is that the separation of the points gives rise to the observed duration profile because, attention can only be shifted from one part of an image to another at a (fixed) limited speed.

These families of results, and others such as Paivio (1971, 1975), show that some mental imagery not only share some of the mechanisms of perception, but also reproduce some of the properties – specifically the temporal profile – of the processes being imagined. The problem for cognitive science is to explain why this should be so.

The possibility that has the least appeal – the theory with the least predictive power – is that this is simply a coincidence. It could be that the variation of subjects response times with the angular separation of the images they are comparing is of no more significance than, say, a particularly high occurrence of even shoe sizes among drivers of green coloured motor cars. It is a logical possibility that it reveals nothing more significant than the fact that testing for large numbers of possible correlations will occasionally find one. In this case, there would be no reason for the correlation, and thus no justification for expecting it to hold in other populations, in other tasks, or even tomorrow. However, while this is a logical possibility, the reliability and diversity of the phenomena involved make it ludicrously unlikely. As a result, any acceptable theory must offer an **explanation** for the observed similarity between the phenomena of imagery and events in the world.

It is the way that some theories attempt to achieve this explanatory power that leads Pylyshyn to suggest that they have the potential to weaken the cognitive architecture. However, pointing out the potential power of a family of theories is far from supporting them. The proponents of these theories believe that imagery shares properties with events in the world as the result of tight constraints upon the possible content and manipulations of the relevant mental information structures. It appears that Pylyshyn's vocation is to show that this belief is misplaced, and arise from a fundamental equivocation in the properties that are ascribed to mental images.

Pylyshyn (1981) opens his argument against the importance of mental imagery by crediting Hebb (1968) with highlighting the distinction between interpreting subjects' reports as describing an image of something with a particular property, or as describing the properties of the image itself.

Reference to distances within an image may be taken to refer to the depicted separation of points on the objects that it contains. But equally, it could refer to the separation of the representations of objects ("In his wedding photograph, Andrew is three inches high"). Maintaining this (admittedly obvious) distinction is crucial when interpreting ascriptions of physical properties, and in particular distances, in discussions of mental images. Unfortunately, doing so is made particularly difficult by the fact (which the arguments below suggest may be inevitable) that people refer to the contents of mental models (or images) in terms of the features in the world that they represent.

In the case of objects in the world, the relationship of distance and time, mediated by the familiar concept of velocity, is a consequence of the necessary laws of physics. Moreover, when discussing things that are **literally** images (in the sense of spatial arrangements of light), then this same relationship clearly carries over. However, this literal interpretation is rarely espoused, and stepping back from the notion of intra-cranial optics sacrifices, or perhaps exorcises, the obvious necessary link between distance and time. If the separation of objects is not represented by distances, then the physics of velocity are quite irrelevant to the manipulation of a mental model, and in particular cannot be used to explain the observed regularity of process durations. It is, Pylyshyn argues, their failure to appreciate this that leads the proponents of mental imagery to grossly overestimate its true predictive power:

> By systematically leaving out the words "representation of" or by using ambiguous descriptions, such as saying that images "preserve relative metrical distances" (which can be interpreted as meaning either that they have or that they represent distances), it is possible to create the illusion of having the explanatory power provided by [the physical laws of motion and velocity] while at the same time avoiding the ontological claim that goes with it (viz., that images are actually laid out in space somewhere inside the brain).
>
> (Pylyshyn 1981. P18-19)

Pylyshyn suggests that there are essentially two ways of explaining the observed correlation between distance ("within" the representation – i.e. separation between represented points) and time. The one that leads to Pylyshyn's recognition of the potential importance of mental imagery – and thus the one he believes its advocates should be taken to be expounding – suggests that the similarity arises from the operation of some physical law which has the same form as that governing motion. Positions and orientations are represented within the brain by some quantity which is literally a physical *analogue*. The alternative is to suggest that the relevant representations are symbolic, their relation to what they represent is arbitrary, and that their manipulation involves a rule-following process which takes a time that is proportional to the distance being represented.

Given Pylyshyn's belief that computationality is a qualitative property, this is a profound gap. An analogue medium could only handle a fixed set of properties, and would represent them in such a manner that they would only allow a very limited range of (non-computational) transformations. Thus the analogue approach offers the possibility of a much weaker theory of cognition because it imposes rigid restrictions on the possible content and processing of representations of the world. In contrast, a symbolic encoding offers arbitrary representational flexibility and is amenable to manipulation by rule-following processes to provide any kind of behaviour at all. It imposes no constraints, and is thus theoretically uninteresting.

Not surprisingly, given the extent of the debate on this topic, there are those who support the causal role of mental imagery who hold that this distinction overlooks an important possibility:

> Pylyshyn's argument loses its force by failing to acknowledge another sense of notions like distance, namely that defined by the inter-cell adjacencies in an array representation and respected by the processes that operate within such an array. According to the theories [of Paivio (1971), Shepard (1981) and Kosslyn(1980, 1983)], position and distance in the array represent position and distance in the world... Thus rather than confusing distance in the world, the internal representation of distance in the world, distance among cells in the internal structure representing the world and distance in the brain, these theorists are making assertions about how these different senses of distance are related.
>
> (Pinker, 1984. P 40, Footnote 5)

Within these theories that use array-like structures, Pinker characterises images as

> patterns of activation in a structure consisting of units (or cells) that represent, by being on or off (or filled or unfilled) the presence or absence of a part or patch of the surface of an object at a particular disposition is space (orientation or location). The medium is structured so that each cell is adjacent to a fixed set of other cells, in such a way that the metric axioms are satisfied.
>
> (ibid)

The theories that use such arrays propose that image transformations occur by repeated small changes. This is most often put forward as the result of the limitations on the connectivity of the array elements, which means that information can only travel long distances as a result of being repeatedly advanced a small step. Thus each cell is continually passing on its contents, which corresponds to a particular portion of object surface, to its neighbour, in what Pinker terms "bucket brigade fashion". Such an account is particularly interesting in that it applies well to both the architecture and ways of using many of the single-instruction multiple-data array processing computers that are currently employed.

In fact, Pinker's criticism of Pylyshyn is not strictly true: he does mention "matrix" theories. His main aim is to show the ambiguity of interpretation of theories specified in terms of computer-language arrays, but does pause briefly to dismiss, literally parenthetically, the possibility of taking such notions seriously:

> (It might be noted in passing that if we were to take the matrix structure seriously, we would be struck with the unavoidable conclusion that mentally represented space is necessarily nonisotropic. This is a formal consequence of the fact that a matrix is a tessellation of some fixed shape and hence has certain nonisotropic properties. For example, if the cells are assumed to be square, then regardless of how fine we make them, scanning diagonally will be faster by a factor of the square root of two than scanning vertically or horizontally. Such an entailment cannot easily be glossed over, except by viewing the matrix as merely a metaphor from some unspecified spatial characteristics)
>
> (Pylyshyn, 1981. P22-23)

This entailment is surprisingly resistant to any attempt to gloss over it in the manner that Pylyshyn's last sentence admits it invites. Nevertheless, it is, of course, possible to defend array theories against such arguments. For instance, it would be possible for an image supporter to avoid predicting any effect of anisotropy by suggesting that the array has enough axes of symmetry that the effect is undetectable. While the geometry of a plane imposes tight constraints on the number of nearest neighbours of tessellated polygons, the relevant property in this case is logical inter-connection, not physical, adjacency. It is not obvious that some scheme of suitably weighted

connections could not be devised to allow each cell to communicate directly with cells along an arbitrary number of axes. Thus while Pylyshyn's point may be troublesome, it hardly seems to be (meant to be taken as) a final rebuttal of such ideas.

Assuming the matter of anisotropy can be dealt with in a satisfactory manner, array theories have much in common with the analogue accounts in which Pylyshyn sees such potential merit. If the inter-cell connections are fixed and provide the only way of transferring information, they can be regarded as simply quantised versions of them. Distance within an imagined situation corresponds to distance along neural paths, quantised in terms of cell connections. Velocity arises naturally from any kind of limitation on either the response time of the cells involved or on the transit time along connections. Thus array theories, like analogue theories, attempt to be predictive by restricting the range of possible manipulations of images to a fixed repertoire provided by the cognitive architecture.

Because they share a single source of predictive power, array theories are also susceptible to the same experimental tests with which Pylyshyn seeks to discredit the truly analogue accounts. Since images are manipulated by the cognitive architecture, which is cognitively impenetrable, imagery phenomena should be equally impenetrable. Thus the experimental results that are used to argue for the role of imagery should be unaffected by cognitive factors, such as the beliefs and desires of the subject or the details of the subject matter of the images. Pylyshyn himself presents a number of experimental results that bring the cognitive penetrability of imagery phenomena into question by apparently altering the observed temporal profile by changing the complexity of the task, or merely the subjects' beliefs about what is required.

As described above, Shepard et al. found that the time subjects take to determine whether two shapes or bodies were identical or mirror images is proportional to the angle between them. Pylyshyn (1979a) presents extensions to this basic experiment that explore the influence of the complexity and inter-relationship of the figures involved. Instead of presenting subjects with either identical of mirror-image polygons, he confronted them with a pair of complex figures, one of which they were instructed to rotate until they could determine whether the other was a true subfigure of it, or a mirror-image of one. He found that the apparent rate of rotation was affected both by the shape of the figure being rotated (quadrilaterals were rotated faster than triangles) and by the sub-figure with which it was to be compared. Thus the rate of rotation of an image is affected both by its content and by other features of the task (i.e. the complexity of the non-rotated figure).

Similarly, Pylyshyn (1981) presents a replication of the work of Kosslyn, Ball and Reiser (1978), and suggests that the correlation of reaction time with distance is an artifact of the subjects' beliefs about what is required in the task. Following Kosslyn et al., subjects were familiarised with a simple map, told to focus on a particular point in an image of it and then asked to determine the bearing of some other point marked on it. He reports that

> When we instructed subjects to imagine a speck moving from the place of initial focus to the second named place, we obtained the same kind of strongly linear relation between distance and reaction time as did Kosslyn et al. When, however, the instructions specified merely that subjects should give the compass bearing of the second place ... there was no relation between distance and reaction time.
>
> (Pylyshyn, 1981. P39)

There was a similar lack of correlation when subjects were instructed to focus upon the second point and determine the bearing of the first.

The fact that subjects can determine interpoint bearings without apparently scanning an image is not new. Kosslyn, Ball and Reiser themselves found the correlation disappeared when subjects were not instructed to use imagery. Indeed, Kosslyn et al. (1979, P537) use this finding to argue that imagery cannot be dismissed as epiphenomenal, because using it affects the temporal profile of the task. However, Pylyshyn's result is crucially different, because in every case his subjects **were** instructed to use imagery (and confirmed in post-test debriefings that they were indeed doing so). Thus it appears that the distance/reaction time relationship is not an essential effect of using imagery.

For Pylyshyn, these results show that both the rotation and scanning of mental images are influenced by the content of the image being manipulated, the purpose of the manipulation and the subject's beliefs about the task requirements. These phenomena constitute a kind of cognitive penetrability that is quite incompatible with the operation of the biologically supported cognitive architecture upon any kind of analogue representation. Given Pylyshyn's views on the absolute nature of computationality, if imagery is not analogue, it can only be computational, and thus just another form of cognitive process.

Obviously, those who ascribe importance to mental imagery reject this conclusion, which means they have to offer other explanations for these awkward results. Thus Kosslyn et al. (1979, P543) propose a mechanism that is clearly applicable to reconciling the idea that scanning images takes a time proportional to the distance with Pylyshyn's discovery that subjects can refocus within an image in a manner that is unaffected by distance. They suggest that subjects also have available a "fundamentally different model of transforming images", which they term the *blink transform.* This allows them to avoid scanning by simply wiping out their existing image (or allowing it to fade) and constructing another which is focussed on the appropriate point. Similarly, they attempt to sidestep the thrust of (Pylyshyn, 1979a) by suggesting that although the overall experimental task is penetrated by the content of the image, this does not necessarily mean that the actual image rotation component is.

> The rotation operation may indeed be a primitive process, but one of the arguments in its "hard-wired instruction format" may be a rate parameter. People may choose in advance slower rates for "worse" probes, perhaps because a serial or capacity-limited process will be monitoring the rotating pattern... They then insert this rate parameter into an appropriate "slot" in the rotation instruction, and rotation ensues.... The important point here is that the determination of the optimal rate is no doubt a penetrable process, but the **rotation** itself need not be.
>
> (Kosslyn, Pinker, Smith and Schwartz, 1979. P546)

Of course, these experimental results are not alone in proving an embarrassment for array-based imagistic theories. For instance, even a simple anecdotal observation is sufficient to prove troublesome:

> When there are parts missing from one's recollections, these are never arbitrary pieces of a visual scene. We do not, for example, recall a scene with some arbitrary segment missing like a torn photograph... When our recollections are vague, it is always in the sense that certain perceptual qualities or attributes are absent or uncertain – not that there are geometrically definable pieces of a picture missing.
>
> (Pylyshyn, 1973. P10)

This forces array theorists to restrict their use to the actual manipulation stage, and postulate that they are stored, and must be "loaded" from, memories organised in an object-oriented manner.[2] Although this suggestion is born out by the experimental evidence of (Kosslyn et al. 1979) that the time taken to form an image depends on the number of items within it, it still represents a considerable restriction on the use of imagery.

Another problem arises from the fact that whenever a body (or image) rotates, the velocity of each part increases with its distance from the axis of rotation. If the cell array is spatially regular, the information describing parts far from the axis of rotation must be traversing more cells per unit time than that pertaining to parts nearer to it. This suggests that the bucket brigades are apparently working faster. This has disturbing consequences for the whole array-based endeavour, since variability in the rate of operation undermines the whole justification for predicting a smooth correlation between distance and process duration.

An alternative explanation, which is prompted by the experimental results described in Finke (1986) is that the resolution of the system falls off towards the periphery, and cells correspond to bigger areas: the brigades are using bigger buckets. However, this is far from a complete solution, since not only does it say nothing about rotation other than about the focus (centre) of the image, but it also raises considerable problems concerning the mechanisms of translation. The intuitively appealing notion of the cells of the array simply passing their contents on to their neighbours must be replaced by a regime where each cell has mechanisms for deciding and providing a suitable (and quite possibly differing) level of detail for each of its neighbours. Alternatively, one could follow Kosslyn et al. and insist that the cell array is spatially regular, but deny that information flow is limited to direct inter-neighbour connection. They argue instead that although the representational capabilities of the system would allow it to employ arbitrary transformations, it persists in using many incremental transitions in order to minimise distortions or to facilitate the extent of the transformation required.

Given the general acceptance that "Mental imagery is a fact one investigates, not a fact one seeks to establish" (Heil: Peer commentary on Kosslyn et al. 1979), this is typical of the style of argument that dominates the imagery literature. Imagery theories are continually being extended to new phenomena or elaborated to cover awkward details of established situations, with the aim of proposing a processing mechanism the operation of which gives rise to the observed mental

---

[2] Note that this argument involves distinguishing weakly equivalent processes by considering features beyond just the behaviour normally produced. As Anderson's mimicry theorem suggests, it would not be possible to say anything about how the information was organised by studying only the "correct" manipulation of images. Any observed behaviour could always be explained in terms of either image-manipulating or proposition-combining processes. However, the consideration of other features – in this case, failure modes – allows these alternative descriptions to be distinguished. It is still possible to conceive of a spatially-oriented ("imagistic") system that could exhibit object-oriented disfunction. However, doing so would involve postulating mechanisms and calculations which had no justification in the normal behaviour of the system. The resulting theory would be cluttered with ad hoc features, and thus clearly inferior to the object-oriented account.

imagery phenomena. However, this natural, and indeed widespread, style of theory development must confront a serious difficulty: Any behaviour whatsoever can be accommodated given enough (freely chosen) parameters, but the resulting theory can only ever describe, and never explain, the phenomena that shaped its development. It must entrust their ultimate explanation to some other theory, by transferring to it the burden of showing why each feature exists and each parameter has the value it has is. As a result, accommodating experimental results in this way is in danger of degenerating into a matter of creating enough parameters to build an engine that can fit the data, with its explanatory power slipping away as each is added.

Not surprisingly, Pylyshyn is at pains to ensure (and point out) that the style of truly analogue theory that he endorses is not susceptible to this style of degeneration. Once a theorist proposes that a particular feature, such as spatial position, is represented by an analogously behaving quantity within the brain, the range of manipulations to which it may be subjected is greatly constrained. This single decision imposes tight constraints upon the cognitive architecture, and thus a whole range of phenomena. Thence arises the explanatory power of the theory. It also commits the theory to enough predictions to enable Pylyshyn to reject it as empirically wrong. Thus Good Science makes progress.

Unfortunately, any proposed "array" architecture can only be claimed to be explanatory to the extent that its performance or structure is determined by factors unrelated to mental imagery. The specification of the rest of its features effectively provides its proponents with an unbounded pool of potential degrees of freedom, and it is disturbing to see how many phenomena seem to require them to dip into it. Thus avoiding anisotropy specifies cell interconnectivity, content altering rotation rate summons forth a suitable parameter and the "blink" transform can only co-exist with any kind of scanning in the narrowest ranges of relative timings or difficulties. Array theorising lacks explicit meta-theoretical criteria. Should an experiment ever reveal that green shapes are rotated faster than other colours on Friday afternoons, only subjective plausibility would oppose having three "primary colour rotation rate" parameters which are normally equal, except after eating fish!

Of course, while a lack of explanatory power is undeniably a problem for a theory, it is not sufficient grounds on which to reject it: **all** theorising begins descriptively. What makes this deficiency fatal is the existence of an alternative account which has both descriptive and explanatory power over a comparable range of phenomena. Pylyshyn suggests that the observed similarity with the world arises not from any constraints on image processing imposed by the cognitive architecture, but from the processes that support it. Of course, these processes could produce any temporal profile at all, which means that the production of the very one involved in the physical process still requires an explanation. The one that Pylyshyn offers is that it arises because the cognitive processes involved are being deployed specifically in order to produce it – that mental imagery is a mode of cognition that involves choosing to simulate events in the world. He suggests that subjects interpret experimental instructions to transform mental images as instructions to form mental images of those transformations occurring in the world. As part of this process, they deploy their (possibly subconscious, or *tacit)* knowledge of the world in order to

produce an accurate simulation of the event, including its temporal aspects. Thus the observations that apparently reveal that mental images are scanned (or rotated) at constant (angular) velocities are simply reflections of the subject's (possibly subconscious) knowledge of the way physical objects move.

This apparently glib suggestion is extremely powerful. It actually predicts, rather than floundering to encompass, Pylyshyn's (1981) observation that changing subjects' instructions disrupts the relationship between their mental imagery and the physical object they are imaging. Moreover, it has explanatory power: on the basis of a single claim – that people tend to simulate the world – it covers a range of phenomena which array theories can only handle on the basis of further assumptions. Indeed, it even offers a principled argument why colour-sensitive rotation rates will not be found – people will not simulate them, because the world is not like that! It also permits obvious extensions to deal with the similarity to the processes of perception by suggesting that subjects deploy their knowledge not of the events being imaged, but of the operation of their own perceptual systems. This would explain why Kosslyn et al. (1979) found the images had finite boundaries which they could overflow, and why Finke (1986) can report that resolution decreases towards the periphery of an image. This is certainly the kind of knowledge one could comfortably hypothesise that people have about their own perceptual processes.

In addition to dealing with many of the "traditional" imagery results, involving subjects' knowledge of the world provides a natural explanation for some imagery phenomena that are otherwise extremely difficult to accommodate. Unsurprisingly, it is Pylyshyn himself who highlights them:

> Imagine a transparent yellow filter and a transparent blue filter side by side. Now imagine slowly superimposing the two filters. What colour do you see in your image through the superimposed filters?... Imagine a transparent plastic bag containing a coloured fluid, being held open with four parallel rods at right angles to the mouth of the bag [i.e. the rods are vertical], and in such a way that the cross section of the bag is a square. Now imagine the rods being moved apart so that, with the plastic bag still tight around them, the rods now give the bag a rectangular cross section. As you imagine this happening does the fluid in the bag rise, fall or stay at the same level (in other words, how does volume vary with changes of cross-sectional shape, perimeter remaining constant)?
>
> (Pylyshyn, 1981. P25)

Thus it appears that inadequate knowledge of the physics of the world can inhibit imagery.

Moreover, what is in some sense the inverse effect can also be observed. Form an image of a smooth-sided object about eighteen inches wide and one foot tall, most of the front face of which is occupied by a transparent plate, from the inside of which shines a changing pattern of light. Now it is obvious how the manipulation of some kind of interconnected array of surface properties can be involved with rotating the image about an axis through the point of observation. However, it is not clear how such mechanisms (or indeed any that only involve suitable descriptions of three-dimensional forms) can alone account for the effect of a rotation about an axis perpendicular to the line-of-sight. The resulting image will almost certainly contain a "back" surface which is either flat or convex and is made of an unpatterned material that is perforated with a pattern of holes or slots. Such an image is related to the original only because of the way television sets are actually made, and hence can only have been generated by reference to knowledge of just that subject.

In Pylyshyn's terms, these two instances of obvious penetration of the imaging process by explicit (cognitive) knowledge makes it very difficult to hold that imagery is in any way a manifestation of the operation of the cognitive architecture, and thus anything other than a form of cognitive processing.

Nevertheless, despite such superiority over array-based accounts, appealing to tacit knowledge remains much weaker than any truly analogue account. The constraints on what an analogue architecture can accomplish would be inviolate, and thus applicable always and to a wide range of tasks. In contrast, the account based on tacit knowledge relies on a cognitive phenomenon, which means its predictions only apply if the subject chooses to make them – i.e. to simulate, rather than use imagery in some other way. Moreover, its predictive power is limited to the extent that the theorist shares his subjects' knowledge of the world. Thus Pylyshyn must regard establishing such an account as something of a Pyrrhic victory. Although it is able to explain a great many of the observed imagery phenomena, it does so in a manner that renders them uninteresting and severely limits their ability to guide theory formation.

However, while the case for tacit knowledge is powerful and well-made, it is not all-encompassing. Firstly, Pylyshyn ascribes many of the effects observed in imagery to subjects' deliberately deciding to simulate the world. This is at odds with the results of Denis and Carfantan (1985), who report that subjects' expressed opinions about the use of imagery reveal that they do not expect scanning (or rotation) to take times proportional to distance (or angle). This is not what one would expect if they were regarding using imagery as the simulation of the physical processes. Moreover, even though a general agreement with the world and the functioning of the perceptual apparatus may be plausibly accounted for by some kind of "subconscious" simulation, some of the detailed observations seem far less likely to be manifestations of subjects' beliefs. For instance, Finke and Kosslyn (1980) found that imagistic resolution, like visual acuity, does not quite have circular symmetry, but falls off slightly more rapidly along a vertical axis. This is certainly a gross mismatch with the the level of detail that most people consciously have about their own perceptual systems.

Any account based on simulation also lacks any natural explanation of the observed interaction of imagery with perceptual processes. For instance, Brooks (1968) reports experiments in which subjects' performance on a mental imagery task is degraded by requiring them to employ visually guided movements to respond (i.e. pointing). Similarly Finke (1986) reports experiments in which the subjects' reaction times in a perception test were influenced by the nature of the mental image they were told to form before the stimulus was presented.

Finally, postulating the cognitive use of tacit knowledge must remain silent on why subjects should spontaneously use imagery in experiments that make no mention of it whatsoever. Finke and Pinker (1982) report an experiment in which subjects were first familiarised with a simple pattern of dots, and then shown an arrow and instructed simply to decide whether the arrow was pointing at any of the dots. Even though the instructions made no mention of imagery, their subjects generally reported that they had made use of an image, and their positive response times were proportional to the arrow to dot distance. Similarly, Johnson-Laird often describes (e.g.

Johnson-Laird, 1975. P13) a simple task that involves describing a collection of snooker balls positioned on a table and asking a simple question about relative viewpoints. Almost everybody admits to using some kind of mental image to answer the question.

These considerations motivate many theorists to argue, contra Pylyshyn, that imagery is **not** merely another kind of general thought. Given its close connection with perception, they naturally express this belief by postulating a special-purpose mechanism with close connections to the perceptual apparatus. Specifically, because their intuitions arose from the prototype of an optical image, they centre their theorising on simulating one using an informationally equivalent representational structure at the interface between the perceptual and cognitive systems. Such an arrangement not only imposes familiar restrictions on the behaviour that could be produced by imagery, but also has a clear connection with the visual system in the two-dimensional pattern of light falling on the retina. However, this last feature of dubious worth. In the peer commentary on their paper presenting their array-based account, Moran offers

> A final comment: the Kosslyn et al. model seems odd from a systems point of view. The visual system extracts visual features, such as edges, from the retinal array and passes these features to the higher-level cognitive system. What seems strange is to postulate another array in which the same kinds of operations, such as edge detection, must be done all over again.
>
> (Moran. Commentary on Kosslyn et al., 1979)

In postulating this kind of dedicated medium and processing engine, array theorists are attempting to embody their constraints within the cognitive architecture, and thus are following, possibly unintentionally, the meta-theoretical guidelines set out in the previous chapter. However, Pylyshyn formulated these guidelines in the light of a qualitative notion of the representational status of a rule which, it has been argued, is not the objective property that his concept of computation needs. This not only undermines his dichotomous view of the computational status of processes, but also invalidates some of the over-restrictive constraints, particularly concerning duration profiles, it leads him to impose upon cognitive theories. Once these notions are rejected, and the idea that cognition is the operation of a flexible autonomous system is introduced, it is clear that this is not the only way advocates of imagery could proceed. For imagery to be a special kind of mental process, and not just another form of cognitive processing, it does not have to be carried out by the fixed processes of the brain, nor involve a specifically impoverished representation. It need only be carried out outside the Flexible Autonomous System that supports cognition.

If this approach provides the opportunity of characterising imagery without insisting that it employ only two-dimensional representation of surface properties, there is considerable reason for taking it. This is provided, in part, by experiments that contribute some of the most persuasive arguments for the importance of mental imagery. Shepard and Metzler (1971) report that subjects found rotation in depth no harder than rotation in the plane of the presented images. As Hinton points out in his commentary on (Kosslyn et al., 1979), this result cannot be accommodated smoothly or convincingly by any account based on a two-dimensional arrangement of information. He himself prefers a definition of mental imagery that Pinker summarises as involving

> information appended to a structural description of the object's shape... [where] ... the various quantitative parameters used in specifying spatial dispositions are encoded as pointers to one-dimensional arrays within which an activated cell represents a particular position, orientation, or size.
>
> (Pinker, 1984. P43)

Note that such a representation of a quantity – one cell within a linearly connected array – is another example of a quantised version of an analogue account. As such, it shares the inability of true analogue accounts to cleanly handle Pylyshyn's evidence of cognitive penetrability.

There is, then, a considerable case for extending the information structures underlying mental imagery beyond simple two-dimensional arrays of properties. In doing so, it is important to recognise that vision is by no means the only sense modality, and has no claim to particular importance beyond an apparent de facto dominance in normally sighted human beings. It is thus wise to consider the extent to which a mental image can contain non-visual information.[3] The desirability of doing so is emphasised by Atrobus who suggests that

> The explanatory power of the [Kosslyn et al. 2D array] theory will be limited to the extent that it is a theory of visual imagery rather than of information processing. No living organism is restricted to information in one sensory modality. Setting this artificial boundary will inevitably limit the kinds of behaviours that the authors attempt to model rather than increasing the power of the model.
>
> (Atrobus, Commentary on Kosslyn et al., 1979, P549-550)

Such a view would find considerable sympathy with many Gibsonians. For instance, Bower (1974) argues that the differentiated senses that are now found throughout the animal kingdom have evolved from a single integrated system. Their apparent diversity masks their underlying unity and disguises their evolutionary role as contributors to a single, unified description of the world. Of course, these multi-modal descriptions are qualitatively richer than any two-dimensional array of surface properties. Since they can contain both sounds and textures, they fit badly with the connotations of calling them mental *images*. Fortunately, there is a term, closely associated with Johnson-Laird, that invokes far more appropriate intuitions. It will be argued that the phenomena of mental imagery should be seen as just some of those that arise from the manipulation, by the cognitive architecture, of a *mental model*.[4] If the two-dimensional array is rejected as the information structure underlying mental imagery in favour or a richer description of the world, then the connection to the perceptual systems must be characterised some other way.[5] This topic provides the focus of the next section.

1.2. What Can Be Perceived?

The task of clarifying the interface between the perceptual systems and the mechanisms responsible for mental imagery is intimately connected with the theory of perception, and the range

---

[3] A matter which amounts to deciding whether there is a sense in which the phenomenon can apply to blind people.

[4] This is not, currently, the most common way to define the term. Certainly Johnson-Laird's (1983. P406) suggestion that "It is therefore safe to assume that a primary source of mental models ... is perception" shows that he does indeed apply it to perception-based representations of the world. However, many theorists, including Johnson-Laird himself, also use the idea of "models in the mind" to describe very different mental structures, many of which are unquestionably subject to cognitive manipulation. This latter use of the term will be discussed below.

[5] As indeed must the range of procedures that can be applied to them, since they can no longer be restricted to being "analogue". This will be dealt with below.

of things that are taken to be perceived. On this matter the most sweeping, and controversial, suggestions are put forward by Gibson (1979). He rejected the idea that perception was the result of processing information about stimuli that happen to fall on the sense organs. Indeed, he objected strenuously to the whole idea that the perception of objects depended on the analysis of other kinds of information, such as descriptions in terms of edges and planes. Instead he argued for complex sense systems that were not just passively stimulated by the world, but actively sought the information they required. These systems provided the perception of objects, and indeed their properties and even potential uses (their *affordances*) in a way that was in some sense *direct*.

Probably the most vehement criticism of this position is (Fodor and Pylyshyn, 1981), who suggest that the most fundamental fault in Gibson's theories is that he fails to recognise the importance of the contribution of individual's beliefs to the perception of the world. Gibson's view is summarised and discussed at some length in Inder (1986). There it is suggested that Gibson's objection arose from a fundamental misconception of the motivation of, and constraints upon, explanations involving homunculi, which simply left him unable to conceive of an acceptable mechanism by which the relevant knowledge **could** be brought to bear. Nonetheless, even Fodor and Pylyshyn acknowledge the importance of Gibson's attempt to broaden the range of properties that can be regarded as perceived. Indeed, the essential spirit of it has clearly influenced Fodor's own ideas on perceptual systems and their effect for the cognitive architecture, which provide the main focus for the rest of this section.

Most psychology involves analysing thoughts and decisions in terms of the operation and interaction of relatively independent mechanisms. Fodor points out that there are two distinct categories of such mechanisms, distinguished by the range of their applicability. Those which have dominated thinking about mind since Plato are widely applicable abilities or mechanisms, often termed *faculties*. However, because Fodor wishes to widen the use of this term, he subcategorises them as *horizontal* faculties, which also have the property that

> The character of mentation is more or less *in*dependent of its subject matter; the faculties are supposed to be invariant from one topic of thought to the next.
>
> (Fodor, 1983. P 11)

Typical of such mechanisms is common suggestion of a faculty of *judgement*.

> Now, this faculty of judgement might get exercised in respect of matters aesthetic, legal, scientific, practical, or moral, and this list is by no means exhaustive. The point is that, according to the horizontal treatment of mental structure, *it is the self-same faculty of judgement every time.*
>
> (ibid. P11 - 12)

The other category of mechanism, which Fodor wants to call *vertical* faculties, was first proposed (in other terminology) by Franz Joseph Gall. Arguing by analogy with instincts, he suggested that different skills were subserved by independent *subject specific* mechanisms. Thus Gall's theory suggests that musical ability is the result of a music faculty responsible for all aspects of perceiving, appreciating and remembering music. Crucially, the similarity of two tunes and the similarity of two faces would be judged by different, task specific, faculties, not by one central (horizontal) faculty of judgement. Moreover, Gall believed that these faculties were specific to certain parts of the brain, and thus were at least partially genetically determined.

Unfortunately,

> Gall made two big mistakes, and they finished him: he believed that the degree of development of a mental organ can be measured by the relative size of the corresponding brain area, and he believed that the skull fits the brain "as a glove fits a hand". Phrenology followed as the night the day, and with it all sorts of fraud and quackery, for none of which Gall was responsible but for much of which he appears to have been retrospectively blamed.
>
> (Fodor, 1983. P23)

Despite this, Fodor believes that the idea of vertical faculties has a vital role in psychology. He supports a specific cognitive architecture which embodies a combination of vertical and horizontal faculties. It has a "central" mechanism which is responsible for "thought" or "general intelligence", whereas perception occurs in a number of independent *input systems*.

Fodor (1983) says very little about the central system, and most of it is arguing that there is little that **can** be said. Among its functions is the formation of an integrated representation of the world from the information delivered by the various input systems. It uses this, in combination with prior knowledge about the world, to drive the fixation of beliefs about the world, which are then employed in thinking, planning and all the other activities of mental life. Note the separation between the descriptions of the world offered by the input systems and the beliefs they engender – a belief about the relative lengths of two lines can be influenced by reading about the Muller-Lyer phenomenon years previously, although the difference in their perceived lengths cannot. Thus, Fodor argues, each of these central processes can involve information from any source, which means that the engine that drives them cannot be subject specific – the central system must involve horizontal faculties. Moreover, and admitting a close similarity to the pessimistic views of Dreyfus (1979), Fodor sees the difficulty of coming to terms with this generality, openness and flexibility as the central problem of cognitive science, and concludes that

> In this respect, cognitive science hasn't even **started**; we are literally no farther advanced than we were in the darkest days of behaviourism.
>
> (Fodor, 1983. P129)

In contrast to the horizontal and incomprehensible nature of the central processes, however, Fodor suggests that the input systems are kinds of vertical faculty. Fodor identifies the human input systems as the traditionally recognised five senses, together with **language**. They play the role (mentioned in the previous chapter) of oracles writing upon the tape of the central system Turing machine, their function being to "so represent the world as to make it accessible to thought" (ibid, P40). As this suggests, their interaction with the central system is limited to the output they produce – the symbols they write on the tape. Thus they constitute a group of independent, autonomous mechanisms, each customised to a particular sphere of activity. Fodor suggests that such functional mechanisms all share a number of distinctive properties which he believes are logically independent – i.e. they need not co-occur. Thus the fact that they do co-occur is empirical evidence that the systems which exhibit them do so in virtue of some common property, and form a "natural kind" which Fodor terms *modules*. However, **logical** independence does not imply **complete** independence, and there are definite conceptual links between the properties he identifies. Nevertheless, enough independence remains to suggest that the notion is a valuable one.

However, the input systems are certainly not merely transducers, in the sense of mechanisms that convert physical energy into neural activity. On the contrary,

> Whereas transducer outputs are most naturally interpreted as specifying the distribution of stimulations at the surfaces (as it were) of the organism, the input systems deliver representations that are most naturally interpreted as characterising the arrangement of **things in the world.**
>
> (Fodor, 1983. P42)

There is an obvious similarity with Gibson's ideas of the direct perception of the distal layout, which Fodor and Pylyshyn (1981) vigorously rejected. The crucial difference, for Fodor, is that the input systems employ complex, possibly experience-based, signal processing to produce a description of the world, and as a result they are free from the need for all kinds of information to be "in the world". Nonetheless, in terms of the overall description of mental functioning, both accounts suggest that the (apparent) layout of objects in the world is simply made available by the perceptual system – "Perceptual analysis is, according to this model, not, strictly speaking, a species of thought" (Fodor, 1983. P43). In short, and as suggested above, Fodor is offering a physicalist's interpretation of the insights offered by Gibson.

Fodor's definition of a module has nine essential properties, for which he argues with varying forcefulness.

(1) Modules are domain specific – they are a type of *vertical* faculty. Thus ability to see and ability to hear arise from quite independent mechanisms. However, this is not to suggest that "sight" is a single module: Fodor stresses that modularity will have finer grain than this, and suggests the possibility of modules for colour perception, voice recognition and sentence parsing.

(2) Their operation is mandatory – they always function, whether they are needed or not. This is born out by the results of Henik and Tzelgon (1982), who find that the comparison of the physical size of two digits is affected by their arithmetic separation, thus showing that the arithmetic properties of the digits are found even though they are not relevant to the task in hand. Similarly, Fodor cites the experimental results of Lackner and Garrett (1973), Corteen and Wood (1972) and Lewis (1970), which indicate the extent to which linguistic processing proceeds even when it is not needed. However, the phenomenon is aptly illustrated in everyday experience:

> "I couldn't help hearing what you said" is one of those clicheès which, often enough, expresses a literal truth; and it is what is **said** that one can't help hearing, not just what is **uttered.**
>
> (ibid, P55)

In many ways, the input systems can be thought of as potentially very complex **reflexes**, a thought captured in Fodor's dedication of the work:

> One day ... Merrill Garrett made what seems to me to be the deepest remark that I have yet heard about the psychological mechanisms that mediate the perception of speech. "What you have to remember about parsing", Merrill said, "is that basically it's a reflex."
>
> (Fodor, 1983)

Just as with reflexes, their continual operation or availability brings survival advantages – no matter how engrossed they are, nobody wants to miss a cry of "look out!"

(3) There is only limited access to the representations that they produce. Fodor suggests that "A plausible first approximation might be that only such representations as constitute the **final** consequences of input processing are ... available" (ibid, P56). The input systems function as "black boxes", inscrutably delivering only their outputs and permitting no access to the means of their production. This, he suggests, is why people who can recognise coins and bank notes, a process which must involve knowledge of their appearance, cannot describe them – the relevant information is inaccessible within an input module.

Fodor goes on to suggest that this restriction is to be expected in input systems. Much of their activity is related to manipulating the welter of information from the sense organs and, to deploy Gibson's terminology, extracting the constancies in the world (Gibson 1979 and copious references there) – e.g. colour and brightness constancy under changing illumination (Day, 1969. Ch. 5), phoneme identification, object permanence under motion (Bower, Broughton and Moore, 1971, Bower, 1982, Chapter 7) etc.

> The **typical** function of the constancies is to engender perceptual similarity in the face of the variability of proximal stimulation. Proximal variation is often misleading; the world is, in general, considerably more stable than are its projections onto the surfaces of transducers. Constancies correct for this, so that in general percepts correspond to distal layouts **better than** proximal stimuli do. But, of course, the work of the constancies would be undone unless the central systems which run behaviour were required largely to ignore the representations which encode **uncorrected** proximal information.
>
> (Fodor, 1983. P60)

The whole "point" of colour constancy is that thought about the world is unaffected – unburdened – by a cloud crossing the Sun, and the myriad changes that this makes to the light falling on the eye.

(4) Input systems are **fast** – very fast. Fodor cites the results of Potter (1975), which show that the visual system can analyse the content of a picture in 150msec, and the results of some shadowing[6] experiments (Marslen-Wilson 1973, 1975) which suggest that the time the brain takes to perform the massive computations involved in understanding speech may be even smaller. Moreover, the time taken is barely affected by the difficulty of the problem – the computational complexity of the problem:

> The difference between a "hard" problem and an "easy" one is measured ... in milliseconds ... It is only in trick cases, of the sorts that psychologists devise in experimental laboratories, that perceptual analysis of an utterance or a visual scene is other than effectively instantaneous. What goes on when you parse a standard psycholinguistic poser like "the horse raced past the barn fell" is, almost certainly **not** the same form of processing that mediates sentence recognition in the normal case. They even **feel** different.
>
> (Fodor, 1983. P63)

The contrast between such processing speeds and the slowness of (conscious) thought and decision making is persuasive evidence for Fodor's suggestion that they arise from qualitatively different underlying mechanisms.

---

[6] Experiments in which subjects are asked to repeat what they hear as they hear it.

(5) Modules are *informationally encapsulated.* They can only make use of a delimited range of information when carrying out their task-specific processing. This greatly increases the possibilities of fast processing, since the problem of even deciding **what** of the system's knowledge is relevant is a formidable task. Thus modules buy speed, which when survival is at stake is a valuable feature, but the cost is that potentially relevant information is inaccessible. Thus "The very same subject who can tell you that the Muller-Lyer arrows are identical in length, who indeed has seen them measured, still finds one looking longer than the other" (ibid, P66). Of course this, and the black box nature of modules, is, as Fodor points out, central to what Pylyshyn calls "cognitive impenetrability" – the operation of the input systems is unaffected by the cognitive system's goals and beliefs. This is precisely what is required of a perceptual system – a successful organism "sees what's there, not what it wants or expects to be there" (ibid, P68).

Fodor admits that there are experimental results that apparently imply cognitively penetrability of perception, but suggests (P82) that "the data look different with a jaundiced eye". He tackles the particular example of the contextual facilitation of word recognition. Samuel (1981) suggests that the meaning of a sentence can affect the speed with which a particular word can be recognised by making it predictable on semantic grounds – a context such as "Because he was afraid of electronic surveillance, the spy carefully searched the room for ..." can facilitate the recognition of words like "bug" and "microphone". However, Fodor cites the results of Swinney (1979), who finds that this same context also facilitates the recognition of "insect" – a word that has no semantic connection with spies or electronics. It is, however, synonymous with (an irrelevant, alternative, meaning of) "bug", and the facilitation seems to be almost a pun on the words involved. Thus, Fodor argues, the perceptual facilitation arises not from the semantic processing of the integrated sentence. It is merely the result of a superficial process that simply reflects the statistics of the observed co-occurrences of lexical items. The mechanism could well involve a form of *spreading activation* (Collins and Loftus, 1975) to explain the essentially similar *phoneme restoration effect* (Warren, 1970). Crucially, this kind of knowledge is domain-specific, of the very sort that could be expected to be held and usefully deployed **inside** a speech perception module. Because of their apparent complexity, modules may disguise their modularity, and the theorist must be prepared to discover it.[7]

(6) Input systems are associated with a fixed neural architecture. Although electronic computers are provably universal, with any given technology the highest processing speeds are obtained by dedicated hardware architecturally customised for the task. Sacrificing flexibility allows more efficient handling of a specific problem, a phenomenon evident from software engineering to evolution, and there is no reason to suppose that the performance of the

[7] Fodor (1983, P119) offers the meta-theoretical observation that it is the informational encapsulation of input modules that makes them comprehensible. He suggests that the limited number of factors that influence their behaviour allows those factors to be identified, and that is why the psychology of perception, in contrast to the psychology of the central processes, is making progress.

architecture of the brain should be any different. However, a problem can only be beneficially handled by a customised sub-system if its demands on the rest of the system are limited and fixed – in other words, if the system is informationally encapsulated. However, once the "hardware" is dedicated to the task, then, as with the analogue computer, there is no penalty in operating it all the time – the mandatory operation mentioned above comes free.

(7) Input modules will have a clear development sequence. This is the most tentative of the proposed properties of modules, although seems a very reasonable consequence of associating a module with a specific part of a growing brain.

(8) Input systems exhibit characteristic and specific breakdown patterns. A particular module will possess a range of possible failures, and their occurrence will produce specific symptoms that will be confined to the functions of the module. Aphasia is a very specific problem that cannot (naturally) be explained in terms of malfunctions of any horizontal faculties like memory or attention. Moreover, there will be a correlation of these breakdowns with physical damage to the brain.

(9) Input systems have shallow outputs. Fodor wants to be clear that they are perceptual systems, not conceptual systems, and tries to limit their abstraction and power. Thus he suggests that deciding whether a remark is ironical involves everything known about the subject and the speakers opinions of it, and as such is certainly not domain specific. This means that recognising irony is done by the (horizontal) central system(s), and not by the language input systems. Indeed, he believes that the language input system performs **no** semantic analysis whatsoever, and the encoding of its output is merely a translation of the sentence perceived into another language. This must, by definition of the purpose of the input systems, be a language interpretable by the central system, and since vocabulary translation requires no analysis of lexical items, it must have a vocabulary virtually as rich as the language being analysed. In short, it is none other than the Language of Thought proposed in (Fodor, 1975). This will be discussed at length below.

Fodor's notion of a perceptual system based on input modules seems to complement the aim of clarifying the relationship between the mechanisms responsible for mental imagery and perception. Thus *mental model* can be taken to be nothing other than a term for a possible pattern of activity at the outputs of the cognitive system's input modules. Of course, this still does not specify precisely what they can represent, although it does constitute a significant step in the right direction. By specifying that models are formulated in terms of the offerings of input modules, it couples decisions about the content of a model to decisions about what can be computed from raw sense data without cognitive intervention. The theorist is still able to postulate that models can represent any kind of information he chooses: only now this is synonymous with a commitment to the belief that there is a mechanism by which information of that kind is **perceived**.

Such a definition certainly accords well with the kinds of things that intuitively seem to belong within mental models seen as generalisations of imagistic theories. It is clearly oriented towards the kinds of properties and relationships typically associated with mental images, such as specific

position (in space), adjacency, orientation, shape, size and colour. In addition, it would include sound, smell, texture and weight, which are highly natural inclusions once the over-emphasis on vision is rejected, while offering a real barrier to those that are more abstract, such as salaries, obligations and favourite pop groups.

However, the power of input modules suggests that many more complex properties should be regarded as perceived than the more traditional notions of perception might lead one to expect. The central argument of Gibson (1979) was based on the notion that sense organs are not constrained to local, static properties, but also detect rates and types of temporal and spatial variation. Thus it is quite possible that input modules detect, and hence mental models can represent, information about object motion (Hubel and Weisel, 1962, 1968) and rhythm, whether in music, speech or events. In addition, since Fodor suggests that language perception is mediated by modules, it follows that models can contain the (some encoding of) linguistic tokens. Similarly, people's sensitivity to highly sophisticated codes for (subconscious) social signaling (Morris, 1977) is equally indicative of the activity of input modules. This would suggest that mental models have the capacity to represent factors such as mood and emotion, as well as inter-personal relationships (pecking orders, attraction etc). Finally, still more surprising aspects of perception are revealed by Michotte's (1963) work with moving dots and other work on the perception of disjointed stimuli, which suggest that the perceptual mechanism is geared towards the detection of (the kinds of correlations that are normally associated with) causality.

The fact that the range of things taken to be perceived, and thus representable within a model, is broader than intuitions based on optical images might predict does not mean that models can represent anything at all. A suggestion that a particular property is represented within a model should be taken to entail a commitment to the existence of modular mechanisms for its perception, which constitutes a real restriction because modules are informationally encapsulated. The identification of such features as careers, attitudes, obligations, and character traits requires the application of unlimited knowledge. As a result, these properties, which are certainly alien to mental image theories, could not be represented within a mental model because they could not be detected by an input module. As a result, perceptibility defines a (somewhat fuzzy) limit on the scope of the perceptual system and the richness of the information structures it can deliver. In exploring how the manipulation of such information structures can be involved in mental imagery, it is essential to examine the kinds of processing to which they can be submitted. This topic is examined in the next section.

1.3. What do Input Modules Output?

In the previous section, it was suggested that the cognitive architecture supports central cognitive processes which are fed descriptions of the world by autonomous input modules. This section focuses upon the interface between these two distinct types of processing.

It is an obvious observation from everyday experience that animals have procedures for predicting the behaviour of the world and directing their immediate actions to suit the situation as it (seems likely that it) is going to be. Most animals see the need to get out from beneath falling objects or in front of approaching cars. Predators head off their prey, which usually have the

foresight to try to escape, and cows gather at milking time. Even the Piagetian *object concept* (Piaget 1937, Bower 1982), which reveals enormous power to predict what will be found in the world, is not confined to human beings (Hunter, 1913).

Prediction of any sort is only possible to the extent that the world is regular, and provided that some mechanism for capturing this regularity is available. Such a mechanism necessarily embodies some information about how the world usually behaves, and all predictive behaviour is the result of the application of such knowledge to the current state of the world. More specifically, it is applied to a description of the perceived world,[8] as it is delivered by the input modules in what might be termed *sensory code*. Unfortunately, knowledge of the physical form of this code, or indeed any other representation used within the brain, remains minimal. However, if one assumes that a code is a consistent means of assigning information content to a physical state, one can characterise it in terms of the kinds of information it **requires** to be specified, the range of features it **allows** to be represented, their **relative availability** and the patterns of loss or confusion between representations couched in it. This is the style of characterisation of a code that is intended throughout the following discussion.

The prediction of the future state of the world being discussed here need not be anything as sophisticated as the ability to explicitly state (or even form an opinion on) what is yet to happen. Indeed, one could quite plausibly deny not only that there is any conscious prediction, but even that there is any consciousness at all, and attribute the behaviour merely to "reflexes". Similarly, describing the system as having "knowledge of the world" makes no claims about the involvement of any form of consciousness. Predictive ability could indeed be learned, with the embodied knowledge being explicitly acquired by the individual animal. But equally it could have grown, with the necessary knowledge innate, the result of generations of experience. In any case it is indisputable that animals, including human beings, are equipped with some means of analysing sense data and producing behaviour that is unmotivated by the immediate state of their environment, but appropriate to a situation that is likely to develop.

The evolutionary use of a such predictive abilities is to protect the creature from threats presented by its surroundings. In a changing environment, this amounts to the cognitive system applying its knowledge of the regularities of the world to the current circumstances, as reported by its sensory organs. A concrete, if sophisticated, example of the operation of this kind of mechanism can be found in the kind of situation typical of those familiar to all lovers of spy films. Just as the hero is making good his escape by creeping past his armed but slumbering captors, he sees his attractive but inept companion accidentally topple a glass decanter from its perch on the shelf above the stone fireplace. He, and indeed the whole audience, immediately recognise the importance of catching it, and in plenty of time for him to do so: the question is how he can do this. Clearly he and they must be able to predict the events to come, and their dire consequences for him, from the information that their senses give them about the situation – that is, its

[8] Notice that there is no need to tangle with the problems that surround matter of whether or not the impulses from the sense organs **represent** the world. It is enough to say that something about them is predictably altered by the state of the world, which must be close to a definition of a sense organ.

description in sensory code.[9] Few people have ever seen a decanter knocked off a mantle-piece beside an armed sleeping captor, and even fewer will have pondered the situation or have been told what happens. It is thus safe to neglect the possibility that the hero simply remembered that letting one fall to the ground was a bad idea. Presumably, therefore, he must in some way have calculated or deduced this fact.

It seems natural to want to tell a story something like the following: Seeing the decanter toppled from its resting place gives rise to some kind of internal description of the situation within our hero's brain. His (possibly innate) knowledge of "how the world works" then interacts with this description, and predicts that, as an unsupported body, the decanter will begin to fall. For the same reasons that he would be unfamiliar with the threat posed by cut glass decanters being knocked from shelves in situations like his, so too will he be unlikely to simply know the dangers associated with their being in free flight, whether or not above stone fireplaces. There is certainly nothing intrinsic to the situation of an object free-falling from several feet that makes it essential to catch it – nobody feels the need to catch the leaves as they fall from trees. The necessity for action only arises when both the nature of the object and the entire situation is taken into account. The situation is still far to complex just to yield to a generalisation, and far too specialised to conceive of him being able to handle it as a "base case" – of just knowing what to do. It must be assumed that he calculates further.

The crucial point is that he must work out that the decanter will indeed fall to the ground. Whether he is seen as doing this by continuing some kind of simulation, or by calculating it from a rule like "unless you have reason to assume otherwise, assume that all freely falling bodies eventually reach the ground" does not matter (although the latter leaves it as a separate matter to explicitly calculate the fact that the decanter is travelling fast). Now, he has arrived at a situation that can be assumed to yield to the application of real-world knowledge. Presumably he recognises the fireplace as a rigid surface and cut glass as brittle, and he knows that when brittle things strike rigid surfaces at speed they break, noisily. Now at last he can see that allowing the decanter to continue to fall will make a noise, and since he is trying to avoid doing just that, it is essential that he catch it. Finally, of course, notice that actually doing so will require him to direct his hand not towards the decanter **is**, but to where it **will** be!

This kind of story seems not only unobjectionable but unavoidable, at least as long as certain details are left to be decided in line with whatever flavour of cognitive theory is most in favour. No claim is made about whether the calculation is an inferential combination of propositions, or an analogue of a manipulation of objects in an image, or even an insightful (homuncular) rational decision based on some kind of description of the situation. Admittedly, it must be accomplished very rapidly, which given the amount of calculation involved places severe (by current digital computer standards) constraints on the computing engine. However, the alternative seems to be to ascribe an enormous amount of a priori knowledge and say that one of the necessary qualifications

[9] Actually, extending "prediction from sense data" to the audience is treating seeing an event in a film to be directly equivalent to seeing the event itself. This is a sweeping, and decidedly implausible, assumption (see below). However, it is harmless for present purposes, since the hero's own ability is sufficient for the point and unproblematical.

for being a super-spy is to know everything that is dangerous in every possible situation. Given that this is undesirable, it is very hard to see how one can support any view of cognition that is not covered by this account! What, then, are the consequences of embracing this uncontroversial suggestion?

The crucial observation is that this story necessarily involves predictions based on the decanter in free fall and striking the hearth. These are entirely imaginary situations – that is, situations represented without external stimulation. This raises the question of whether such representations are in the same code as actually perceived situations. Explanations of the spy's mental activity can be characterised by whether or not the imaginary scenes are represented in sensory code.

This clean distinction is clouded by the work of Marr (1982). This indicates that (at least visual) sensory representations will pass through several different representations as it progresses through the perceptual system. However, this does not matter if the inter-code translation problem is considered with the model of the analogue computer in mind. This suggests that the brain can be seen as having wired-in, or dedicated, translation circuits, so that all the levels of Marr's analyses are always available without cost, because their encodings are what might be called *freely translatable*. This means that if any internally generated representation is introduced into the system in **any** of the codes that externally generated representations pass through, it too would be immediately available in all subsequent ones.[10] Therefore, as far as all processing that occurs after this stage is concerned, the two sources of information are equivalent.[11] Thus the possibility of freely (inter-) translatable representations forces a re-formulation of the characterisation of systems. The important distinction is whether internally generated situations are held in (not **the** format but) **any** of the formats freely usable for externally generated information, as opposed to some other, distinct, format.

The matter of whether "imaginary" situations are represented in the same way as the senses report the environment is further complicated by the fact that perception and thought are not the only ways to become aware of a situation and its potential consequences. They can also be readily appreciated merely by understanding a verbal description of the situation, whether in a novel or a thesis, as is clearly illustrate by the fact that the hero's plight was immediately obvious from the above description. Once again this must be the result of prediction, brought about by the application of world knowledge to some kind of representation of a situation. However, in this

[10] This discussion is phrased in terms that assume there is a clear direction of flow of information. This holds in many theories of perception, including, of course, Marr's. However, it is possible to imagine systems involving feedback, where some stage of processing is affected by (something in the same format as) its output. If both translation processes are suitably supported, the codes are freely inter-translatable. The system will simply appear to be using a single code which exhibits functional characteristics which are blend of those of the "component" codes (at least to the extent that there is evidence to suggest it is meaningful to discuss them).

[11] There is also the possibility of an analogous problem relating to storage format, rather than encoding. This is illustrated by the familiar digital computer, where the same information may appear in patterns of capacitor charges or transistor activity in main memory, but in the alignment of magnetic particles in the backing store (where it may also be interspersed with chaining pointers and checksums). These, too, are freely translatable, since the (potentially complex) process is transparently carried out by dedicated hardware, so the main argument goes through. Moreover, as mentioned above, the the physical form of mental information structures is never described, primarily because nobody has the faintest idea how to start. As a result, the possibility that it is actually flitting between many different ones is, for the clearly foreseeable future at least, of no consequence.

case the internal representation of the situation did not arise from its perception, but from its (English) description.[12] Explanations of this phenomenon, as with descriptions of the spy's own behaviour, can be characterised on the basis of whether or not situations appreciated as a result of understanding descriptions are encoded in the same way, in sensory code, as those that are perceived.

The most parsimonious suggestion is obviously that, regardless of whether it is perceived, imagined or heard about, the representation of a situation will be couched in the same code as the results of perception. However, it is by no means logically necessary, and parsimony alone is hardly compelling. Therefore, it is perhaps worth briefly exploring some of the consequences of rejecting it and assuming that linguistic input, recollection or imagination result in representations that are phrased in some other, *internal* code that is distinct from any of those that are used for direct sensory perceptions.

At this point it is important to recognise the important consequence of the formality condition. This implies that a computational system can access nothing but the physical form of its data, which means that the system can only perform the function it does on the basis of fundamental assumptions concerning what and how those data represent. In particular, the neural machinery responsible for carrying out the predictive processes can operate a representation only on the basis of its physical instantiation. This means that the (possibly innate) knowledge involved, and maybe even the mechanism for applying it, must also be instantiated or encoded in line with the same code as the representations to which it is to be applied. Unfortunately, this fact makes attempting to maintain a real distinction between sensory and internal code representations problematical.

One area of difficulty relates to the development of the predictive mechanism, and the problem of **checking** any prediction (couched in internal code) against the subsequent state of the world (presented in sensory code). But there are further difficulties even within the prediction process. When the spy reacts to his predicament, the prediction must be made on the basis of real (sensory code) information – that actual sight of the decanter being toppled from its perch. But in the case of verbal descriptions or the subsequent manipulations of the results of the initial predictions, the representations involved are in internal code. Internal code is defined as a format which is not one that external data is normally translated into, so by definition the sensory data is not translated into it for analysis. Thus the predictions cannot all have been made by the same mechanism, since they were made on the basis of different data. Moreover, one cannot avoid this by equipping a common world-predicting mechanism with two separate "front ends", one for each code. This would only work if both translated into a common representation on which the mechanisms operated, and since the mechanisms would normally be applied to sensory information, this would be a code that sensory information would normally be translated into. However, this is a contradiction, because internally generated representations would be being translated into, and thus presented in, a code normally used for sensory data.

---

[12] Given the number and importance of conventions – the use of music, cutting between simultaneous events, flash-backs etc. – to the interpreting of films, it is arguable that understanding what is happening on a cinema screen has as much in common with language as with simple perception.

Attempting to maintain the distinction between internal and sensory codes thus forces the recognition of the existence and independence of two separate sets of world knowledge – one for each of the codes. But this too has problems. What is the motivation for postulating such a double system? Occam's Razor puts the onus firmly on those who would support a distinct internal code to argue for its existence – to justify why they believe that one should regard imagined situations as being represented and processed differently from the situation detected by the sense organs.

A potential source of support for such a notion would arise if the two sets of procedures were not identical, thus giving rise to detectable failures to transfer knowledge between experienced and reported situations. This might give rise to situations where a normal adult would notice no anomaly in a story of a two-year-old toddler lifting a car off the ground. Alternatively, there might be knowledge about unexperienced situations that could not be applied to the real world. In some cases one might wish to say that this is indeed the case – most people are quite familiar with the idea of a Dalek but would still be very alarmed and confused to meet one. But such "failures" can be explained in many other ways which would try to capture this as a sensitivity to the fictional nature of the Lords of Skaro. More convincing would be failures to transfer information that really ought to be transferred, such as inability to **act** upon instructions such as "if any of the lights on this panel comes on, put this plug in the socket below it" despite being able to correctly say what should be done when asked to **imagine** such a situation. The very peculiarity of these suggested problems emphasises that they are, to say the least, not commonly observed features of everyday life.

The alternative is to accept that the two sets of world knowledge are exactly equivalent, and seek other kinds of observable behavioural effects that might distinguish dual from single systems. The conditions that seem most likely to produce observable consequence involve attempting to tackle two or more related tasks simultaneously. Under the twin system theory, the ability to deal with the real world should not be impaired by contemplating an imaginary scene. Under a single-code theory, there might well be interference, at least on the assumption (plausible on the basis of experience with actual computers) that it is difficult to apply the same knowledge or procedures to two sets of data at once. Unfortunately for those who would support the twin systems of representation, introspection reveals that it is very much easier to imagine a situation with closed eyes or in a calm environment, and there is ample experimental work that supports it (e.g. (Brooks, 1968)). Thus the evidence, such as it is, supports the single-code approach.

In conclusion, if imaginary situations are thought not represented in the same way as the world as perceived by the senses, one is obliged to postulate the existence of a large amount of "duplicate" knowledge for interpreting situations in each of the two forms. Occam's Razor suggests that one should not take such a step without compelling evidence which everyday experience at least suggests does not exist. Hence anyone proposing theories of the operation of mind should be happy with the idea that imaginary situations can be thought of as represented in the same code/format as externally perceived situations. In other words, part of the brain is to be seen as manipulating internally (re)generated information in packets just like those it receives from the sensory apparatus.

Accepting this observation now allows us to make progress with the concept of a mental model. Imagine a line drawn across the path of sense information at the point just before it is translated into a format that is (freely translatable into one) used for internal representations. The processing and translation carried out before that line can be thought of as the perceptual mechanisms of the cognitive system. If these are treated as a "black box" system, they define a mapping from objects in the world to "data structures" (or at least, patterns of neuronal activity) within the brain. What is more, since these structures are encoded in the same way as internally generated situation representations, the inverse of this same mapping can be used to define the (set of) states of the world that the internal representation **corresponds to**< The existence of this mapping allows the possibility of describing a shared-code representations in terms of the corresponding state of the world. Thus one can refer to a mental model of a green pyramid atop a blue cube without the slightest notion of how, or even **which** particular features are explicitly represented.

Crucially, it has been suggested that this representation is in a form that is used for both perceptual and internally generated representations. Moreover, it is amenable to manipulation in line with any accumulated knowledge of the behaviour of the world. How the manipulation of such information structures could form the basis of mental imagery will be elaborated in the next chapter.

CHAPTER 2

Models and Modules Beyond Perception

2.1. Other Modules

The later sections of the previous chapter argued for a cognitive architecture in which input modules deliver an integrated, multi-modal description of the world, dubbed a *mental model*. This section attempts to suggest how these ideas can serve the objective proposed at the end of the first section, to establish a richer characterisation of mental imagery.

Although the main thrust of Fodor's arguments is directed towards establishing the usefulness of the concept of an **input** module, he also muses briefly on the possibility that there might also be *output* modules. This is a very natural suggestion within a computational theory of mind. Folk psychology, which expresses its predictions as personal-level generalisations, is very successful. The computational approach attempts to square the success of its predictions with the physically arbitrary nature of its terms by suggesting that a description of the world in terms of the relevant properties is (physically) present within the brain, and is manipulated in line with suitably encoded rules. In such a context, input modules provide the initial translation from physical to personal-level properties, but this is only half the story. Folk psychology employs not only personal level properties, but equally intentional descriptions of actions. Pylyshyn (1980, P161) pointed out the impossibility of explaining why people commence building evacuation behaviour in terms of the physical properties of their environments, but the behaviour itself is equally incapable of description in terms of the contractions of leg muscles. The relevant generalisations not only involve the safety of remaining in the building, but also what behaviour is appropriate to leaving it.[1] There is a clear need for a means to translate between personal-level intentions and muscle activity. To the extent that this is the inverse of the function of the input modules, it is obvious to consider whether the mechanisms involved are equally modular.

A moment's thought shows that movement (or muscle control in general) possess many of the features of modularity. It certainly involves very rapid processing of task-specific information, as is shown by the control problems confronted in robotics. It is associated with specific neural structures with characteristic breakdown patterns (e.g. paralysis induced by strokes) and undeniably exhibits a development pattern – toddlers do indeed toddle. However, concerning the properties of mandatory operation and informational encapsulation of modules, it is much harder to decide

[1] Or, indeed, raising the alarm or getting others out of the building, which may even involve going **into** it, despite having full knowledge of the danger involved.

whether they are exhibited, or even in what sense they can be expected to be applicable.

While the requirement that the operation of input modules is "mandatory" has a straightforward interpretation, it is far from clear how the term should be applied in the context of controlling behaviour. A sense organ can supply signals to several processing systems simultaneously, but a limb can only perform one action at a time. Thus while input modules can operate concurrently, no system of output modules could function without effective coordination of their activity. Cognitive processes serve to provide this control by deciding what behaviour is most suited to the circumstances, and must therefore be able to ensure that the most appropriate actions, and no others, are actually taken. It is thus clearly inappropriate to assess whether output modules are mandatory in terms of inability to prevent their activity – it is no surprise that healthy people can easily inhibit all their "output modules" simply by deciding to keep still! Admittedly there is some activity, even involving voluntary muscles, that people cannot suppress, but to think of output modules as exemplified by such processes as the blink reflex would be to ignore the spirit of Fodor's message about the complexity of modules.[2]

This problem can be tackled by suggesting that input modules are described as mandatory to capture the fact that the cognitive system can neither prevent or circumvent their operation. This means they are autonomous, since although the cognitive system can **ignore** the descriptions of the world that they produce, it cannot affect or prevent their production (Fodor, 1983. P53). Similarly they are obligatory, because the descriptions they produce are the **only** access to the world that cognitive processes can have. Subdividing the mandatory nature of modules in this way reveals a sense in which it can be applied to the mechanisms that regulate behaviour. Although they must be in some sense controllable, to allow cognition to do its job, they can nonetheless still be seen as mandatory to the extent that they are autonomous and unavoidable.

If the mechanisms responsible for producing behaviour have these properties, this will have observable consequences. If taking action can involve only invoking output modules without any control over their internal operations, then what modules are available will restrict the range of possible behaviour. There will be actions that people simply cannot (not merely do not) carry out, even though they want to. Skills are obvious examples of this kind of inability,[3] made particularly striking by the fact that even when acquired they are not normally transferable between hands. Even more striking, and entirely free from possible objections concerned with timing or not knowing what is required, is the widespread inability of people to wiggle their ears, roll their tongues, raise just a single eyebrow or emulate Mr. Spock's Vulcan salute.

Similarly, many aspects of some behaviour strongly suggest the involvement of autonomous mechanisms. Among the most pervasive of these is posture control. People do not simply collapse when they stop thinking about remaining erect, even though doing so requires the precise coordination of innumerable muscles. Walking over uneven ground or grasping something involve still greater complexity, yet both are apparently effortless and literally thoughtless. Moreover, the

[2] Of course, the blink reflex can only be thought of as simple to the extent that the detection of something moving towards the eyes is simple, a suggestion that owes its plausibility to Gibson.

[3] Skills and their acquisition will be discussed further below.

mechanisms involved cannot be dismissed as merely servo systems triggering muscles in line with crude sense data. People preparing to grasp and carry objects implicitly take into account not only weight and strength distributions, but also subsequent activities that should not be impaired. Similarly, everyone continually and subconsciously adjusts their posture, both to avoid fatigue and as a means of social signalling (Morris, 1977).

Much of this, such as inter personal distance, eye contact and even pupil diameter, is sufficiently subtle to be normally imperceptible. However, gesticulation while speaking often varies between obvious and distracting, and even on occasion becomes laughable. Where it is important to control a speaker's image, gesticulation can be suppressed, although (initially at least) doing so requires real conscious effort. Interestingly, it is often accomplished by the speaker actively concentrating on some other task that involves the hands, such as grasping the rostrum. The cognitive decision to prevent gesticulation is most readily carried out not by direct inhibition, but by invoking another module to override it. This has obvious connections with Fodor's observation about the difficulty of not taking taking note of the offerings of input modules, where

> In the interesting cases – where this is achieved without deactivating a transducer (e.g. by sticking your fingers in your ears) – the strategy that works best is rather tortuous: one avoids attending to x by deciding to concentrate on y, thereby taking advantage of the difficulty of concentrating at more than one thing at a time.
>
> (Fodor, 1983. P53)

The other problem in requiring that the properties of modularity apply to the mechanisms controlling behaviour relates to the flow of information. Fodor argues that module boundaries restrict the flow of information both into and out of the module. Thus cognitive processes are unable to access the intermediate results of a module, while the state (beliefs) of the cognitive system are equally inaccessible to the module.

It is easy to show that there are restrictions on information flow involved in the production of behaviour. The ability to throw something to a target requires accurate and sophisticated use of the laws that govern projectiles. Nevertheless, any school physics teacher will testify to the difficulty of instilling a conscious appreciation of these laws, and McCloskey (1983) shows that many people complete their education without acquiring one. His work also suggests that the longevity of Aristotelean physics is a phylogenetic precursor of this ontogenetic phenomenon. Presumably people living prior to the early sixteenth century could play catch, yet that ability seems not to have interfered with their universal acceptance of a system of beliefs based on the idea, illustrated by diagrams such as that reproduced in (Burke, 1985. P138), that all terrestrial motion was rectilinear. Similarly, an ability to pronounce "bad" and "pad" distinctly is rarely accompanied by even the faintest notion of the concept of voice onset time.

Unfortunately, while one can cleanly delimit the information employed by an input module, the analogous characterisation of output processes is problematical. Obviously, if they are to direct activity in line with the (personal level) objectives of the cognitive system, output modules must have access to those objectives, but the above examples of output module behaviour show that this is only a small part of their information requirements. Even the blink reflex only triggers when there is something moving towards the eyes, while the environmental information required by the more complex output modules suggested above is obviously much greater. Walking requires a

suitable route and information about the state of the ground, grasping is affected by beliefs about the weight of the object and what is to be done with it, and posture and gesticulation are shaped by social circumstances and communicative intent. Thus all these modules are sensitive both to circumstances and to information that is mediated by cognitive activity. Nevertheless, although the information requirements of output modules clearly cannot be characterised as tidily as those of input modules, there is no reason to believe that they are unconstrained.

The fact that output modules require information about the state of the world immediately raises the question of where they obtain it, and in particular whether it is mediated by the operation of input modules. Although these were characterised as being informationally encapsulated and mandatory, these properties were only argued for with respect to the cognitive system. It would thus be perfectly possible to suggest that other processes within the brain, such as output modules, interact with the world by completely independent processing of the deliverances of the sense organs. However, one of Fodor's more powerful arguments for motivating the "black box" nature of input modules is that the results of their processing describe the world more accurately than what they were derived from. It would thus seem reasonable that output modules can be expected to use the descriptions of the world produced by input modules, because doing so enables them to function more effectively.

Once the concept of a module has been extended beyond the initial interpretation of sense data, it suggests a way of overcoming a serious flaw in Fodor's arguments. One of the most obvious arguments against the idea of "black box" modules is that certain information that is undeniably part of the module's functioning is **sometimes** accessible to general thought. Thus while Fodor can illustrate the closure of modules with the unavailability of the details of syntax for normal cognitive processing ("Which did I just say...? Was it syntactic details or details of syntax?" (P57)), there is no doubt that people **can** have access to a range of such surface phenomena if they require. Fodor stresses that this phenomenon might well be related to a failure of memory, but is still forced to suggest that the information is only ever available at all because the modules are "leaky" – opaque, rather than black, boxes.

However, the impenetrability of modules can be retained in spite of their apparent "leakiness", thus offering the potentially greater explanatory power, by suggesting that they can be *chained* – that certain modules work off the output of others. This allows the availability of certain intermediate results, but preserves the mandatory and impenetrable nature of the overall process, which suggest its modularity, since each process involved is itself a module. However, if the later modules in the chain are doing anything worthwhile, and if they are not it is hard to justify postulating them, their output will be more useful than their inputs, the intermediate results – the meaning of a sentence will generally reward attention more than its syntactic structure. As a result the latter will be attended to and the former (normally) ignored and, therefore, not committed (retrievably) to memory.

If chained modules are postulated to be directly (i.e. permanently) coupled, the resulting system has all and only the properties of a module with a specific "leak". However, radically different properties appear when the inter-module interface is flexible, even penetrable. Under

these circumstances the later modules in the chain can be invoked independently. Surprisingly, perhaps, this does not impugn their impenetrability or opacity – the ability to operate a calculator is quite independent of the ability to affect, or indeed knowledge of, the way it behaves. The result is what can best be termed a *processing* module.

To illustrate how processing modules can result in a cleaner system, consider language processing. The discussion will focus on its perception, which Fodor offers as an example of a modular input process, although precisely analogous arguments apply to its production. There are two main problems with treating language perception as a single unified input module. The first is the fact that, as illustrated above, people can on occasion have access to "partially perceived" language, such as syllables, which surely belong within the inaccessible interior of the language module. Fodor notices this problem, and it prompts him to weaken the definition of modularity and admit that modules are not entirely impenetrable. The second he does not mention at all: people can read (and even lip-read) as well as hear. Yet, if modules are simply informationally encapsulated processors of sense data, it is not clear how a language module that deals with speech can respond to stimulation of the retina.

However, a far more satisfying description of this situation results from accepting the idea that language perception is the result of a **processing** module. It normally operates on a description of linguistic material provided by another (input) module, which work on speech perception suggests could well be in terms of syllables. Crucially, there are at least two such modules, one each for extracting linguistic information (syllables) from speech and from text. Like all other modules, they both produce their results (report the detection of syllables) in "internal code". This explains not only how a single language module can respond to this information in a modality independent manner, but also how it can be available to other, possibly cognitive, processes. Of course, information about what syllables have been heard or seen is of very limited use, apart from for supporting the perception of language, and there will be very few processes that will take advantage of this availability outside psychology laboratories. However, a great deal of other information that plays a role in language perception has important other uses: tone of voice not only indicates syntactic groupings, but directly reveals a great deal about the speaker – gender, emotional state etc. The free availability of this kind of information can be explained without relaxing the restrictions of information flow across module boundaries by recognising language interpretation as the result of a processing module working on information encoded in internal code.

These arguments have suggested that the concept of modularity that Fodor brings to perceptual processes can in fact be observed equally clearly in other areas of mental activity. Almost all of the characteristics that he ascribes to a module can be observed in both the production of behaviour and internal processing. The most notable exception concerns the use of dedicated "hardware" (or "wetware") to support the functions. This is obviously necessary for the earliest stages of perception and the last stages of motor control – the relevant processing simply has to occur on the end of the relevant signal-carrying nerves. However, its applicability to processing modules is less obvious. There is indeed considerable evidence that a specific area of the brain (Zurif and

Blumstein, 1978) can be identified with syntactic analysis of language, although the evidence for associating any other mental operations with any particular area of the brain, while less clear cut, is nonetheless persuasive

Gazzaniga and LeDoux (1978, P133) describe the recovery of a keen gardener from a severe stroke. During this process, he passed through a period during which he was conscious and spoke normally, but had no access to most of his knowledge of gardening. During this period, he could identify something as a flower, but could offer no more detail and accepted its identification as a carnation as a piece of new information. He also recalled clearly that he had recently planted something with his daughter, and even where he had planted it, but had no idea what it was he had planted. These anomalies passed as the condition subsided and his recovery progressed. This complete loss of a specific area of expertise because of disruption to the function of an area of brain tissue is strongly suggestive of the kind of localised *vertical faculty* that is being proposed.

Similarly, the results of Stamm (1969) from experiments involving the application of electrical stimulation in order to disrupt the functioning of a region of the brain also show the localisation of specific cognitive functions. In the experiment, a monkey was allowed to see a morsel of food being hidden in one of two wells. An opaque screen was then interposed for a period of 8 seconds, at the end of which the animal was allowed to retrieve the food. Previous experiments involving its surgical destruction have shown that a particular part of the animal's neocortex plays a vital role in the animal's ability to perform this task. Stamm hoped that the use of electrical stimulation to temporarily disable the structure would allow him to determine the precise stage in which it was involved. He found that the animal's ability to retrieve the reward from the correct well was dependent (only) on the unimpaired functioning of the relevant area of the neocortex at the time the food was hidden. These results show the localisation of one particular function – namely noting the continued existence of the food at the time it is hidden – in an area of the brain which is not involved in its subsequent retrieval.[4]

In addition, it is possible to suggest that the model itself may be similarly associated with a neurological structure – specifically, with the *corpus callosum*, the bundle of nerve fibres that links the two hemispheres of the brain. When this structure is destroyed, much of the capacity to communicate between the two sides of the brain is lost. Indeed, because of the way sense information is distributed between the hemispheres, it is possible to train and test each side of the brain separately. One side of the brain may, through its control of parts of the body, exhibit skills and knowledge of which the other side is completely naive. Such experiments show that following their surgical separation, both halves of the bisected brain behave as an intelligent cognitive system (although as a rule only one hemisphere – normally the left – has significant linguistic abilities). Finally, it is noteworthy that severing the corpus callosum is used, as a measure of last resort, to alleviate the symptoms of otherwise intractable very severe epilepsy. Such attacks bring about complete loss of control of the body and obliteration of all mental processes – precisely the kind of

[4] Or alternatively, the desirability of investigating a certain well is noted by an area of the brain that is not involved in subsequently doing so.

global effect that would expected to be associated with a disorder of such a central feature of the cognitive architecture.

Unfortunately, these observations are still very coarse. If processing modules are as specialised as "knowledge of flowers", it seems likely that there are a great many of them distributed somewhere within the "associative cortex". Moreover, the hope of physically finding them is lessened by the possibility that the operation of a number of independent modules can be supported by a single physical structure. This is certainly a coherent possibility, as the existence of modern time-shared computers demonstrates. Moreover, what little is known about connectionist architectures suggests that they are very well suited to supporting this kind of poly-functionality, though without the performance limitations that intuitions fostered by the sequential processing engines might predict. Fortunately, dropping specialised neural structures from the requirements of modularity serves to weaken its importance only very slightly. The essential identification of modules is based on their **functional** properties, and the current state of neurology means that it can contribute little in terms of how these may be supported. Nevertheless, analysing mental phenomena in terms of functional modules still has consequences for the (functional) properties of the rest of the architecture. This topic is developed further in the next section.

2.2. Implications for the Cognitive Architecture

Fodor's *Modularity of Mind* proposes a cognitive architecture in which many self-contained input modules autonomously offer their descriptions of the world for the attention of a central "computational" process which underlies cognition. The previous section argued for the extension of these ideas to include output and processing modules. However, doing so requires much more of the cognitive architecture than simply an increased number and diversity of modules. If (any part of) a module is restricted to manipulating information from just a single source, it is possible to envisage it being tightly and permanently coupled to that source. However, if modules are able to respond to information from a number of sources – such as syllables heard or read – the architecture must provide mechanisms for selecting and distributing the relevant information. Similarly, if modules are to be able to request specific information or actions (particularly relating to the visual system), the architecture must direct the necessary communications and ensure their intelligibility. Thus the cognitive architecture begins to take on the form of a mechanism for regulating the intercommunication of a number of autonomous processing modules.

In recent years Computer Science, in the guise of Artificial Intelligence and under the banner of "expert systems", has devised and begun to study a computational architecture with just these properties. Originally used in speech perception (Lesser et al., 1977 and Erman et al. 1980), *blackboard systems* are now being recognised as a generally applicable computational architecture with highly desirable properties. The essential style of interaction it was intended to support is based on that between a number of experts cooperating on a problem. The architecture's name arises from the idea that they are gathered round a blackboard, on which is described the problem to be solved. Progress towards a solution is achieved as the experts each add any contributions they can make to the blackboard. This may be to identify a sub-problem to be solved, to offer a solution to one, or to point out a difficulty with a contribution made by another expert.

In an ideal case, the blackboard would be the only medium of communication between the experts, and its operation obviously depends on the various experts adopting conventions for placing information on the blackboard. Not least, there must be a means for deciding which expert should be allowed to write when, particularly where contradictory views or intentions arise. In addition, there must be agreement on the interpretation of what is actually written, for instance by adopting a single universally applicable coding convention so that every expert could understand, act upon or criticise everything written.[5]

It should be clear how this kind of description applies naturally to the cognitive architecture Fodor proposes, and the suggested extensions to it. Input modules can be seen as observers, continually updating a description of the current environment that they have written on the blackboard. As a result, the state of the blackboard constitutes an integrated, multi-modality description of a state of the world in a format that is available for general processing – in short, a mental model. Within this framework, cognitive processes function as one of the "experts", assigned the task of monitoring the state of the world, or model, and deciding what behaviour is appropriate. Typically, it will concentrate on the most high-level descriptions that any modules can offer, and initiate any desired action by adding a similarly high-level description to the blackboard. Finally, output modules will be continuously scanning the blackboard for messages that pertain to the kind of activity they control, on receipt of which they will initiate the necessary muscle actions.

However, cognition would not be the only module reading and amending the model. So too would other processing modules be continually seeking the kinds of information that they can process, and signalling their results by adding them to the blackboard. The prediction of the behaviour of the world, discussed in the context of recognising the consequences of a tumbling decanter, constitutes an obvious example of such activity. Moreover, as was suggested in that discussion, the usefulness of these modules is greatly increased because they can treat perceived and imagined situations alike, an ability that relies on the fact that the input modules themselves are not the only source of descriptions in the formats that they produce. The blackboard may contain either an actual description of the world itself, or a similarly formatted description created internally. In either case, processing modules are able to manipulate and amend the model it presents, quite independently of the Flexible Autonomous System supporting cognition. But, of course, such autonomous processing of sensory-code descriptions of the world is nothing less than the characterisation of imagery proposed at the end of Section 1.1!

The architecture presented, then, allows a characterisation of imagery as a non-cognitive form of processing without having to involve any kind of planar array or quantised-analogue component. As such, its special status is in no way dependent on the **form** of any representation, thus satisfying Anderson's worries about the indeterminacy of such matters. Perhaps surprisingly, however, it also enjoys the principle benefits of Pylyshyn's account. As was pointed out above, the relevant

[5] In fact, universal intelligibility is less advantageous than it may appear, since each expert will typically only deal with information relevant to a few specialities or tasks, and simply ignore everything else. All that is actually needed is to ensure that information of a given type is always written in the same format, and somewhere it doesn't interfere with other, unrelated, information. This means that any expert using a particular kind of information effectively need only be able to accept it from one source.

processing modules can only predict the behaviour of the world because they are able to deploy knowledge of how things usually behave. Moreover, even though this knowledge may well have been acquired from experience of the world, it can still be tacit. Indeed, it may well be even more tacit than Pylyshyn's monolithic notion of the application of computational ideas to the brain may let him conceive. Even though it may well take the form of "rules" to follow, it may be nothing whatsoever to do with cognition. Those rules may have been written by, and for the benefit of, a completely independent flexible autonomous system, within which they reside, totally and forever inaccessible to any cognitive process.

This account, which is possible once it is recognised that the cognitive architecture can be decoupled from the functions provided by the structure of the brain, can naturally account for many imagistic phenomena. The autonomy and potential specialisation of processing modules clearly have the potential to explain why they should be more effective at model manipulation than the more general purpose cognitive system. As a result, it is natural to predict that it will be advantageous to utilise them whenever possible, which amounts to suggesting that imagery will be spontaneously adopted even without any prompting from experimental instructions. Moreover, the fact that imagery shares with the perceptual processes the use of a single "blackboard" provides a point of contact which is able to explain both the constructive (e.g. Finke, 1986) and destructive (e.g. Brooks, 1968) interactions between them.

Equally, however, the essential knowledge-based nature of processing modules also allows very direct explanations of many of the restrictions on imaging ability. This obviously applies to the examples of colour overlap and bag of fluid cited above from (Pylyshyn, 1981). It also explains why rotating an image of a television seen from the front brings into view the back of a set: experience of real televisions has taught that this is what happens. Moreover, this also suggests a natural explanation of why images being rotated pass through all intermediate points. Continuous rotation is literally an everyday observation, and the ability to survive in the physical world will have motivated the development of efficient procedures for predicting it. In contrast, direct transformation from one orientation to another is a rare phenomenon to say the least, and there is no reason (and very little assistance) for any processing module to have mastered the art of dealing with it. Note that such an explanation makes no appeal to the difficulty of direct transformations, nor any inherent restrictions on the representational medium that prevent it. It doesn't happen in imagery simply because it doesn't happen in the world.

Finally, in passing, it is perhaps worth mentioning that Fodor himself, who generally disparages the role of mental imagery, should have no objections to regarding his proposed cognitive architecture in this light. Although he may never have considered the kind of direct inter-module communication it suggests, it presents no more challenges to his approach than the influence of cooking smells on salivation. Its description meshes perfectly well with the enterprise that he has been championing since (Fodor, 1975), once it is accepted that what is on the blackboard is written in the Language of Thought. Cognitive processes can still be seen as the manipulation of such sentences in arbitrarily complex ways, precisely as Fodor has always characterised them. All that has changed is that they have lost their monopoly in this respect.

Processing modules are also empowered to make such "inferences", although only within a restricted domain – they, too, speak the Language of Thought, only their vocabulary is somewhat limited.

2.3. Other Processing Modules

The previous section concluded that Pylyshyn was indeed correct to point out that at least some manipulation modules are knowledge-intensive processes. Much of this knowledge could be ascribed to the genetically specified structure of the brain reflecting the evolutionary advantages that accrue from the ability to predict the world. However, at least some – e.g. rotating a television set – is clearly learnt. This suggests that it is important to consider the way in which other manifestations of memory fit within the proposed blackboard architecture. Doing so leads naturally to the suggestion that memory is supported by a modular system.

To see this, consider answering the question "Did you hear about a man biting a dog?". This surely does **not** involve recalling every incident related to a man, searching among them to select those that have to do with dogs, and then finally rejecting those that do not involve biting.[6] The number of events relevant to each individual concept is far too large to be sensibly moved about at semiconductor speeds, let alone by neurons. Moreover, even "pointers" to the information involved (if such a concept is meaningful) would be orders of magnitude greater than even the most generous notions for the capacity of "working memory". It simply makes no sense to speak of retrieving everything to do with any one of these concepts – even if long-term memory had the bandwidth to supply the information, there is not even a hint of a mechanism that could do anything with it!

The answer, obviously, is that long-term memory itself is responsible for, and capable of, selecting only those events that involve **all** the concepts required. The *spreading activation* account pattern matching, as embedded in the architecture of Fahlman (1979), is one of the most influential proposed mechanisms for this kind of retrieval. It involves assigning a "salience" value to each memory element according to its closeness to the one of the concepts of interest, then adjusting these values in line with the relationship to each subsequent concept, and then finally taking account of all the events that have accumulated a certain score (or possibly just considering the highest single score).

Of course, this process must not only take account of the isolated concepts, but also the relations between them – none of the myriad tales of poodles harassing postmen have any relevance to an attempt to recall a situation involving a man biting a dog. The sequential consideration of "concepts" just described is arguably the conceptually simplest or most obvious approach, and also, and quite possibly not unconnected with this, accords well with the operation of von Neumann

[6] This question undeniably has the false ring of a linguists example. Natural processing, and thus better intuitions, might be sparked by "Do you remember when Sinclair had some kind of a fight with a guy from Acorn?". Fight might refer to anything from a battle in World War II through to the world heavyweight final. Fights involving Sinclair might be anything from a game on a Spectrum computer to a playground scrap over the ownership of a C5 or the merits of programmable egg timers. Even making Sinclair one of the combatants leaves for consideration all kinds of struggles for market share, even if the other is related to Acorn.

computing engines. However, what is known about the brain gives no reason to expect it to proceed in this way. Quite to the contrary, its architecture of many slow but heavily interconnected processing elements strongly suggests parallel operation, which appears to simplify handling the relations between constituents of the query.

This argument leads to a characterisation of one (though quite possibly not the only) kind of access that long-term memory can support. It is able to supply information in response to a simultaneously presented combination of concepts (entities, properties or whatever) and the (possibly complex) relations holding between them. However, such a specification of the information to be retrieved is nothing less than a partial description of the event of interest. Thus the suggestion is that memory can be seen as a processing module which is able to respond to an incomplete description of a situation on the "blackboard" by filling in further details the specific event that is most compatible with it. Such a mode of operation has much in common with a very natural style of interface that is commonplace within data-base management systems, known as *query by example* (Zloof, 1977). Such interfaces are very natural, and allow access to portions of large and complex data bases without requiring any reference to how the information is stored or any special concepts or notations.

This last feature is particularly relevant because postulating such methods of memory access does not require any additional representational capabilities of the cognitive architecture. All that is needed is the ability to represent possibly incomplete descriptions of things, situations or events – in short, what has already been characterised as a mental model. Furthermore, it does not require the module that is requesting information to have any knowledge of how, or even where, it is stored, which in line with the informational encapsulation of the processing module(s) that support the process. Finally, it accords well with the known properties of "connectionist" networks, which will settle in the nearest stored stable state nearest the state used to interrogate drive or probe them (see Hopfield, 1982). This allows them to exhibit *reconstructive* behaviour, since they will stabilise in (i.e. retrieve) a stored pattern on the basis of being driven, or probed, by only a part of it. The process is precisely akin to a ball bearing coming to rest in the bottom of a bowl, having been placed somewhere within it.

Having suggested that episodic memory is effectively a modular process, it is natural to consider the extent to which other kinds of learning can be characterised in terms of the behaviour of modules. One obvious example is apparent from the fact that skilled behaviour has already been identified with the operation of processing and output modules. This means that the fact that people can acquire a new skill indicates an ability to develop new (or significantly enhance existing) modules. Similar arguments go through in terms of the specialised perception associated with (the appreciation of) certain kinds of expertise. Novice ball players are often handicapped by the fact that in some sense they simply do not see the spin on a ball, while competent musicians appreciate much more of the structure of the music they hear than those who are less trained. Indeed, both these aspects come together in the (not uncommon) ability to learn a new language.

Since skilled behaviour involves the operation of processing modules, studying it can provide insight into the way in which modules interact. Because the knowledge it requires resides within

an impenetrable module, it cannot be directly accessed in any sense, but must instead be elicited by indirect probing. Indeed, this inaccessibility of the rules of everyday grammar was one of Fodor's principle arguments in favour of their impenetrability, and the linguists' standard methodology constitutes just such probing. Processing modules, such as that supporting syntax analysis, operate by scanning the blackboard for information of the kind they can deal with and where possible adding to it the results of their specialised processing. The linguist tries to fathom how English syntax is handled by putting example sentences on the blackboard and noting the analysis module's response. Indeed, the same approach seems to be adopted in answering questions concerning the detailed steps in any skill. Very few people who can tie a necktie immediately know how many times they grasp and release it in the process. When trying to answer this, they typically **imagine** themselves to be in a tie tying situation, a process which involves suitably configuring the blackboard. This provokes a processing module somewhere to divulge its knowledge, by adding to it a series of instructions concerning what to do next, and the question is answered by counting the relevant graspings as this sequence unfolds.

The same kind of interaction can be observed in the way that skills are acquired, which is in many ways the reverse of this process, in that the relevant knowledge must be transferred **into** a module. Since the module is informationally encapsulated, this cannot be done by simply "showing" it the relevant information. Indeed, since modules are vertical faculties, no other module would be able to deal with it anyway. The cognitive system may know that a tennis backhand begins with "taking the racquet well back", but this is not at all the level of detail or kind of information that will help a processing module dealing with muscle coordination. Once again more indirect approaches are called for. Because it can come from no other source, the module must acquire the detailed knowledge it needs for itself. This explains the overwhelming importance of **practice**, and why no amount of instruction can ever replace it. However, it is possible to improve the efficacy of practice, and the precise techniques that are effective for this reveals something of the limitations of modular systems communicating via a blackboard.

Consider the way skills are typically conveyed.[7] The first stage, and always one of the most effective techniques, is to have the learner simply watch examples of the required behaviour and attempt to copy it. Watching the skill being executed has the effect of transferring to the blackboard, and thus bringing to the attention of any "interested" processing module, at least an outline specification of what is required. The same kind of phenomenon can also be observed in children, who undoubtedly acquire many of their behavioural traits from their parents, and it has been argued (Meltzoff and Moore, 1977) that the necessary mechanisms are both innate and observable in neonates only seventeen minutes old. Moreover, such mimicry is not confined to humans: parrots are famous for it, and the saying "monkey see, monkey do" is not entirely unwarranted. The fact that seeing an action is sufficient to allow it to be copied is at least revealing about the kind of information that output modules can accept. It is also more than

[7] The example will be couched in terms of physical (motor) skills, but similar characterisations also hold for perceptual and intellectual skills.

somewhat suggestive of a close connection between the blackboard representation of the creature's own activities and those of others.

Once the student can at least approximate the behaviour with some consistency, the instructor will attempt to make improvements. Very often, and as a natural extension of the initial process, this will take the form of the coach demonstrating once more and highlighting, often by exaggerating, a particular feature. Equally, the student might be given exercises which incidentally involve a (possibly exaggerated) form of the required behaviour – such as swinging a tennis racquet with a book in the head cover, in order to develop follow-through. Alternatively the instructor might physically guide the students hand. In either case, the stated objective is to ensure that the student literally "gets the feel" of the action, which in effect amounts to presenting the relevant processing module with a highly detailed description of what is required. Finally, the instructor might liken the required behaviour to some other activity – tennis coaches often suggest the correct service action involves "scratching your back" and "throwing the racquet at the ball". Once again the objective is to get a detailed specification to the processing module, this time by assuming that it is already contained within the specified actions.

Finally, this approach is able to explain an otherwise puzzling phenomenon. The processing modules responsible for skilled behaviour operate by accepting from the blackboard a description of a situation and a (cognitively) desired objective, and contributing to it a specification of the appropriate actions. However, the output modules responsible for actually executing these actions can be prevented from actually doing so – for instance, by giving them overriding commands to remain motionless. This makes it possible to exercise or rehearse the processing modules simply by simulating an appropriate situation on the blackboard and monitoring the actions they suggest to predict their probable consequences. This predicted behaviour can then be used to provide feedback to the processing modules, and provided the prediction is based upon accurate knowledge of the world, the net result will be to improve its preformance. Of course, this kind of manipulation of the state of the blackboard is nothing less than mental imagery. This leads to the startling prediction that simply imagining performing a skill may well enhance the ability to actually execute it. However, Mendoza and Wichman (1978) and Ryan and Simons (1982) have presented experiments which suggest that this is indeed the case. Indeed, this is a procedure advocated by many sports coaches.

While the above discussion has dealt with skill development, the form of learning that has attracted most attention, at least at the boundary between psychology and philosophy, relates to concept acquisition. While concepts must obviously arise as the result of experience, their acquisition is generally regarded as a qualitatively different process from merely remembering relevant instances. Instead, it involves extracting the appropriate generalisations, allowing novel instances of the concept to be recognised and handled. The relation between these generalisations and the individual instances that gave rise to it is clearly similar to that between the memory of a particularly sweet tennis shot and a knowledge of how to volley. There is clearly a close connection between skill and concept acquisition, which suggests regarding a concept as a combination of a skill in recognising something as an instance of that concept, and a skill in

knowing the implications of this. Note in particular that this emphasises the gulf between knowing of a concept, or learning the definition of a term, and being able to use it – it is as wide as the gulf between reading a book on skiing and surviving a difficult piste.

Indeed, not only does conscious knowledge of what is required not give rise to the correct performance, it often has quite the opposite effect. The immediate result of suggesting how to improve a skill is usually a deterioration in its execution – telling someone what they are doing wrong just before an important performance is inevitably a disaster. The processing modules responsible for the skilled behaviour simply cannot assimilate information directly from the cognitive system, and any conscious attempt to shape behaviour serves primarily to interfere with the smooth execution of the skill in its established form. The result is the familiar problem of "trying too hard", and advanced training in so many skills relates to striking the right balance between effort and relaxation. Martial arts training in particular is heavily oriented towards preventing (conscious) uncertainties and distractions from degrading performance, and urges its practitioners to perform without "trying". Indeed, the notion that superlative performance comes not through consciously trying, but by acting with an "empty" mind is central to Zen philosophy. Similarly, one of the biggest obstacles to skiing well is the tension caused by the steepness of the slope triggering instinctive responses that cannot be suppressed. In such cases, expertise critically depends on preventing irrelevant activities, and one of the hall-marks of an expert is the effortless of their excellence.

While ascribing both skill and concept acquisition to the same general mechanism might be aesthetically rewarding, it nonetheless begs the question of how either of them is actually achieved. However, there is evidence from experiments in connectionist architectures (see the papers in Hinton and Anderson, 1981) that such systems can exhibit something vaguely akin to this behaviour. Certainly the kinds of systems discussed by Hopfield (1982) are able to settle into stable states which correspond roughly to a number of stored stimuli, but precisely to none – that is, they are capable of some kind of "generalisation". In addition, Hinton has described multi-level networks that have been able to learn complex mappings by using intermediate nodes to detect useful sub-classes of situations. Even though the kinds of situations being handled are undeniably far too restricted to grace with the term "concept", they do at least suggest that the kind of mechanism they embody might be able to provide the right kind of behaviour if implemented with the scale and complexity of the human memory. When combined with the ability to regenerate a complete pattern from a partial specification, they offer the possibility of conceiving of a system that will support a smooth transition from episodic memory to concept memory, with both being handled by essentially the same query by example mechanism which treats specific events as simply highly specific concepts.

Within the cognitive architecture, it would be possible to suggest that processing modules arise from the differentiation within some kind of connectionist memory system, with both specific instances and elaborate "concepts" being able to update the blackboard using the same, uniform mechanism. Note, however, that such a connectionist system can hardly be described as involving the execution of explicitly represented rules – its operation would not constitute "computation"

under the definition of Fodor and Pylyshyn. Nevertheless, since it is able to modify the internal weightings that direct its behaviour, it still constitutes a Flexible Autonomous System. Moreover, until research provides a suitable formalism for describing the operation of very complex examples of such system, computation provides the only conceptual tool available.

Having made this suggestion concerning acquiring new concepts, it is important to clarify its relation to Fodor's (1975) lengthy argument for the logical impossibility of doing so. Given his logico-linguistic leanings, he expresses this in terms of adding a new predicate to a language (of thought). He focuses solely on the identification of exemplars of a concept (as opposed to elaborating the implications of having done so), which he suggests is the very least that is required to learn it. This necessitates formulating an expression which is co-extensive with the predicate to be learned, which is not always possible, as he illustrates with the impossibility of capturing the meaning of the universal quantifier of predicate calculus in propositional logic. This leads him to conclude that

> What has been argued is, in effect, this: If the mechanism of concept learning is the projection and confirmation of hypotheses (and what else **could** it be), then there is a sense in which there can be no such thing as a new concept... To put it succinctly, the concept-learning task cannot coherently be interpreted as a task in which concepts are learned. Since, barring rote memorisation, "concept learning" is the only sort of learning for which psychology offers us a model, it is probably fair to say that if there is such a process as learning a new concept, no one has the slightest idea of what it might be like.
>
> (Fodor, 1975. P95-96)

Fodor is, of course, right to suggest that the power of the representational system determines the range of concepts that can be expressed. As a result, human thought must indeed be underpinned by a representational system that is able to represent every concept that human beings can ever acquire. However, it is only by spending too long among the logicians' infinite sets of possible worlds that this restriction can be seen as imposing significant limitations on mental abilities. To see this it is only necessary to recognise that, since it can store (for present purposes effectively) unlimited amounts of arbitrary information, such as the complete Encyclopaedia Britannica, a VAX computer can in some sense represent almost any human concept. But it would still be ludicrous to say that it in any way **has** those concepts.

What is wrong is that, even in those aspects of learning a concept that relate to recognising an exemplar, there is much more than simply being able to express the relevant conditions. In particular, they must be made applicable to the world as it is experienced. This entails that processing resources must be dedicated both to refining the conditions for its recognition and to the accretion and interpretation of information applicable to its instances. No finite system can pursue these objectives for every possible combination of representable conditions – that is, every possible combination of combinations of low-level representational primitives. Adding a concept to a cognitive system provides the seed for the crystalisation of this information. In this task the concept terms of natural language do indeed have a role to play, not merely as Fodor (1975, P85) suggests as a form of abbreviation that allows more complex ideas to fit within some kind of processing constraint. Instead, they act as markers for positions in the space of possible concepts that others have found useful.

Finally, the fact that skills are encapsulated within processing modules allows a natural explanation for another puzzling experimental observation. DeGroot (1965, 1966) has found that after a five second exposure to a position from a game of chess, strong chess players are able to recall the positions of the pieces very much better than other people. However, where the position they study is simply a random arrangement of pieces, this effect disappears. On the theory proposed, the expertise that the chess player has build up over years of practice resides largely within specialised processing modules. When he studies a board position, these modules will recognise the important relationships between the pieces, such as attacks and pins, and add them to the blackboard. This means that in the case of a "sensible" position, such as from a real game, it will be a much richer representation of the board that is committed to memory, which will increase both the potential cues accessing it and the redundancy for checking what is retrieved.

Notice that on such an account, the master can be literally said to *perceive* a threat. It is added to his mental representation of the situation by precisely analogous mechanisms to information about the content of speech or the spin on a ball. In the same way, an experienced accountant can simply "see" a cash flow problem and a car mechanic can "hear" a failing prop shaft.

2.4. Constraining the Architecture

The first chapter identified the primary objective of cognitive theorising as the definition of a functional architecture that would serve to provide the basic operations that subserved all cognitive functions. The argument so far has been directed towards establishing the credibility of the architecture known within Artificial Intelligence as a *blackboard system.* As proposed, it owes much to Fodor's *Modularity of Mind*, but extends his notion of a *module* beyond *input* modules, which relate to the initial processing of sense input, to include both *output* and *processing* modules. It was proposed that these modules can communicate only by exchanging information via a single universally accessible information structure representing a (possible) state of the world which was termed a *mental model.* Within such an architecture, it was suggested that mental imagery can be seen as the manipulation of the model by a processing module that is outwith the flexible autonomous system responsible for cognition. Such an arrangement was shown to be capable of capturing many of the most influential phenomena and intuitions on both sides of the imagery debate.

Chapter 1 also pointed out that when attempting to characterise the computational architecture, sound methodology dictates seeking to postulate the weakest possible mechanism. Any proposed theory must obviously have sufficient flexibility to be capable of producing the whole range of observed behaviour in an interesting range of tasks. However, to be anything more than descriptive, it must aim to be able to produce that behaviour in the fewest ways possible. The more distinct ways in which a theory can accommodate to a particular observation, the weaker (less falsifiable) it is. Conversely, the greater the number of ways a system can handle a situation, the less possible it is to predict which will actually give rise to observed behaviour.

In this respect, the least desirable theory is that proposed by Fodor (1975), which essentially ascribes the more complex mental operations to an engine which is able to appropriately

manipulate knowledge expressed in a "language of thought". Fodor does suggest that some features of the syntax of this language may influence mental processing. However, he essentially imposes no upper limit upon the complexity of its constructs or the extent of its vocabulary – indeed, quite the contrary, he argues (pp. 124-156) that they must be comparable to those found in English. As a result, this account imposes no restrictions on the range of representations or processes upon them. Essentially, Fodor is arguing that human mental abilities, particularly those related to language, are so intricate and subtly inter-related that they can only be seen as arising from a single, monolithic mechanism. This would be an innate processing agent, able to act upon sentences in the Language of Thought in a manner appropriate to their content within the context of other similar sentences expressing beliefs and desires pertinent to the situation being considered: in short, a homunculus.

Of course, it may be that Fodor is right and the only characterisation of human mentation available to our current conceptual apparatus is as the appropriate manipulation of belief sentences by a mechanism that is both innate and inscrutable. But to suggest that this variation on folk psychology may be the best that can be achieved is surely the council of despair. Obviously, every theory of human mentation must propose a system that is able to understand the subtleties of human mental life. While it may be that this can only be accounted for as the deliberations of a homunculus, psychologists must at least strive to gain some insight into the nature of its internal machinations. Given the current profound ignorance of how the brain processes information, this must necessarily be done in a functional manner. The best that can be done is to attempt to describe the ability – to analyse the homunculus – in terms of the lesser, more readily understood, components and the interaction between them. This is, of course, precisely what the blackboard architecture amounts to.

The desire to (eventually) accommodate the entire range of human mental activity **may** require postulating a mechanism which can both represent and manipulate even the most subtle, abstract concepts of human thought.[8] If so, weakening the architecture entails limiting as much as possible both the power of its techniques for doing so and the range of situations in which they are applicable. Within a blackboard architecture, the most sweeping curtailment of flexibility is achieved by restricting the communication between modules, which amounts to limiting the representational capacity of the blackboard itself. This can impair not only the availability of problem-related information for specialised manipulation, but also meta-level information for controlling processing. When applying an existing theory to a new task or observation, it is important to resist the temptation to do so by adding an appropriate representational or processing mechanism to the proposed cognitive architecture. This is the road of the rotation rate parameter, and it leads to an account that is ad hoc and at best merely descriptive. The explanatory power resides in the **in**abilities of the architecture that the theory carries over between domains.

---

[8] The alternative being to seek an account in which these are in some sense "emergent" abilities, arising, like the power of an orchestra, from the interaction of a number of lesser contributors. Individual neurons do not have beliefs.

The discussion thus far has suggested that the manipulation of the state of the blackboard can be identified with mental imagery. In doing so, a mental model has been characterised as a representation of a situation in a form that might have been produced by the sensory systems. While this hardly defines the range of representable properties, it does at least highlight an area that is relevant to constraining it. In phylogenetic terms, the model developed to allow the manipulation of the knowledge of the environment gathered by the sensory systems. As a result, the kinds of thing that can be perceived will have shaped the range of information that the code in which it is represented can accommodate, while the exigencies of survival in a hostile environment will have honed its organisation. As a result, phenomena observed in the study of perception can, potentially at least, have a bearing on the content and organisation of the information structures present on the blackboard. This is not at all as tight a restriction as the traditional notions of imagery might lead one to expect, and the richness of the phenomena that input modules allow to be characterised as "perceived" loosens it still further, but it may well be the best that can be achieved.

Even though they may be much looser than desired, it is possible to pursue the objective of imposing limitations on inter-module communication by requiring that it be based on a mental model that can contain only perceptible properties. Trivially, since it is represented by a finite mechanism, there are limits on the capacity of a model. More significantly, the previously mentioned work on the effect of response modality on imaging ability by Brooks (1968) gives results showing that imagery performance is degraded by even very simple additional visual aspects of the task. This suggests that there are severe limitations on the ability of the mechanisms underlying imagery to deal with more than one image (or model) at a time.

The suggestion of a restriction on the ability to model multiple distinct situations actually gets support from one of the standard anti-imagery arguments which Fodor cites.

> Consider the Tiger and his Stripes... If seeing or imaging is having a mental image, then the image of the tiger **must** – obeying the rules of images in general – reveal a definite number of stripes showing, and one should be able to pin this down with such questions as "more than ten?", "less than twenty?". If, however, seeing or imagining has a descriptional character, the questions need have no definite answer... "numerous stripes" may well be all the description says.
>
> (Dennett, 1969. P136-137)

Since he admits that imagery appears to represent a significant type of mental processing, Fodor (1975, P188-191) indicates a number of counters to this argument. Unsurprisingly, the one he prefers, which is based on the idea of images under description, serves primarily to emphasise the fundamentally propositional nature of the processes that underly mentation.

Fortunately, there is an alternative explanation. It is a fact of everyday experience that the perception of a collection of things is not accompanied by an awareness of precisely how many there are. This is not the result of any indefiniteness in the perceptual process, merely that total numbers of entities is not one of the perceived features of a situation. There is, in fact, good reason for this, since in most circumstances there are an enormous number of (sub) sets of things it would be possible, and utterly futile, to count. In order to find out how many objects have a certain property, it is necessary to decide precisely how to recognise such things and then to

explicitly **count** them. Since it is supported by mental models, which only contain the kinds of information about a situation that is offered by the perceptual processes, the same situation must hold in the case of mental imagery. The image of the tiger may have a determinate number of stripes, but discovering how many that is will involve counting them.

Counting, of course, will involve individuating each stripe in some way which, given that images have finite resolution, may well be impossible given an image of a complete tiger. This can be solved by "zooming in" – invoking a processing module to supply the missing detail, and create a close-up view of a small enough number of stripes to be countable. Since counting stripes in a small patch of tiger does not reveal its total stripe inventory, it is necessary to maintain the count while scanning the enlarged section along its entire length. This process obviously requires access to the context of a representation of the complete tiger. Unfortunately, if it is impossible to entertain more than one model at any time, that original image had to be destroyed to make way for the "close-up". The model of the tiger may have a definite number of stripes, but they can't be counted.

In a similar manner, the response of the perceptual system to ambiguous or noisy sense data can be used to support severe limitations on the representation of uncertainty. For instance, each edge of a Necker cube is always either at the front or the back: never both at once, and never neither, either! A given picture is either an old lady or a young girl, a duck or a rabbit. In each of these cases the perceptual system offers only one of the possible interpretations of an ambiguous stimulus.

> The mapping of layouts onto effective stimuli is certainly many-to-one, for it has been repeatedly shown in psychological laboratories that percepts can be caused by samples of the ambient medium which demonstrably underdetermine the corresponding layout. Nor is this phenomenon specific to vision. Consider, for example, the *phoneme restoration effect* (Warren, 1970) in psycholinguistics.
>
> (Fodor and Pylyshyn, 1981. P72)

In each of these cases, the perceptual apparatus offers only a single interpretation of the available data which in no way reflects the other potential interpretations. Indeed, a similar tale can be told for the case of the so-called Garden Path sentences, in terms of a language processor committing to one interpretation of a locally ambiguous language stream, and thus being unable provide the correct interpretation when disambiguating information becomes available.

The fact that the perceptual apparatus exhibits this kind of behaviour should be unsurprising. Every detector system, including the human sense organs, suffers from "noise" – random fluctuations in the physical environment which will lead to spurious features in its output. Clearly whatever system makes use of the detector output must have some means of dealing with this uncertain data, and one possibility is to consider only the single most likely interpretation. The principal alternative is to explicitly represent the uncertainty involved and make subsequent reasoning sensitive to a range of possibilities. Indeed, this approach can be seen in many expert systems that reason with uncertain data (See Alty and Cooms, 1984. Ch. 5), such as MYCIN (Shortliffe and Buchanan, 1976) and PROSPECTOR (Duda et al., 1976). As a result, the matter of which (or, indeed, what **mixture**) of these is adopted constitutes a fairly constrained but necessary parameter of a cognitive theory.

It is tempting to argue against an approach based on explicit manipulation of uncertainties, both in terms of the additional computation it seems to require and the increased complexities of learning from indeterminate situations. However, the actual difficulty of dealing with these matters cannot be assessed in the prevailing absence of a clear-cut idea of the low-level operations of the cognitive architecture. As a result, the choice between the two approaches remains a free parameter. What can be claimed, however, is that the evidence offered above suggests that as a matter of empirical fact the cognitive system operates on the basis of commitment to the perception of the most likely state of affairs. This brings with it justification for a commitment to the notion that the mental model is precise and unambiguous. Interestingly, these are properties which Pylyshyn has suggested are a key characteristic of mental images (and thus, by extension, models) that allow them to be discriminated from propositional information structures:

> It would be quite permissible... to have a [propositional] mental representation of two objects with a relation between them such as "beside". Such a representation need not contain a more specific spatial relation such as "to the left of" or "to the right of". It would seem an unreasonable use of the word "image" to speak of an image of two objects side by side, without the relation between them being either "to the left of" or "to the right of".
>
> (Pylyshyn, 1973. P11)

Indeed, this is distinction is widely endorsed by Johnson-Laird, the principal proponent of models.

Finally, the perceptual system is necessarily restricted in its ability to handle negative information. For every entity in the world, there are an infinite number of things that it is "not". This information clearly cannot be represented in any (necessarily finite) description – the best that could possibly be achieved would be a highly selective indication of the absence of certain properties. Moreover, this selection could only be made by a highly knowledge-intensive process that could compare the detected properties of an object with those it could be expected to have on the basis of all the available, contextually relevant knowledge. Such an operation is, of course, absolutely the antithesis of the informationally-encapsulated processes that Fodor characterised as modular. In resolving this conflict, it is important to bear in mind the obvious fact of everyday perception (and, indeed, prototypical models): it always captures what things **are**, never what they are **not**. Even where the actual object fits the description of "a car with no wheels", what is actually **seen** (or modelled) is a car propped up on bricks (or a hovercraft, or a tank or whatever). In such a situation, the methodological objective of postulating the strongest possible constraints on the cognitive architecture suggests proposing that the perceptual system can detect, and thus the mental model can represent, **no negative information whatsoever.**

The arguments presented so far have been directed towards proposing that a particular functional architecture, based on a number of independent, task-specific Flexible Autonomous Systems, underlies all mental activity. Given such a proposal, the methodological objective of maximising the number of falsifiable predictions that a theory can make requires weakening this architecture as much as possible. This has been done by suggesting that inter-module communication is confined to the contents of a single information-structure, termed a "mental model". The non-cognitive transformation of such a model has been suggested to give rise to mental imagery. Moreover, the phenomena of imagery, particularly in relation to perception (or the

simulation thereof), has been used to motivate the suggestion that the model is only able to represent perceptible properties. Unfortunately, the mediation of perception by powerful input modules means that this encompasses more than might otherwise be expected. Nonetheless, it has been suggested that a mental model is only capable of representing one single description, which can contain neither ambiguity nor negation.

The universal applicability of the architecture means that these restrictions, derived from mental imagery in perception-related tasks, should have implications for other areas of psychology. Johnson-Laird, one of the most vigorous proponents of model-based theorising, focuses on two main areas to which the theory can be applied. The first, which is hardly surprising considering the extent to which it dominates cognitive science, is language. The next chapter discusses what is possibly the most obvious extension, namely the identification of the proposed mental model with the popular idea of a *discourse model*. The second concerns its possible implications for general (abstract) thought, which Johnson-Laird exemplifies by syllogistic reasoning, a topic which provides the focus of the second half of this thesis.

CHAPTER 3

Models, Language and Thought

3.1. Introduction

The suggestion that the functional architecture of the brain constitutes a collection of independent task-specific Flexible Autonomous Systems has been developed in the context of perceptual and perception related phenomena. However, it is the nature of an architecture that it underlies all mental activity, and features postulated on the basis of imagery can be expected to have implications in other areas. The rest of this work deals with these implications. The first section of this chapter focuses on language, and specifically comprehension, and proposes the identification of the popular idea of a discourse model with the contents of the blackboard. Subsequent chapters deal with abstract thought processes, as exemplified by the categorial syllogism. However, suggesting the relevance of models for such purposes, at least as it was put forward in it older guise of mental imagery, has provoked many arguments which claim to show that the role such mechanisms can play is severely limited. As a result, the final sections of this chapter are devoted to the consideration of such anti-imagery arguments. At least some of them will be dismissed, either because they apply only to simplistic, ''two-dimensional array'' forms of image theory or because they are the result of a view of mental processing that has been overly influenced by language.

3.2. Models and Language

Theories of language comprehension that involve the construction of a representation of the situation being discussed currently have considerable popularity. One of the most wide-ranging is that proposed by Johnson-Laird, which will form the focus of the next section. However, it has clear ties with other accounts, although they are not all so explicitly expressed in terms of the modification of information structures. Work essentially on formal semantics led Kamp (1981) to suggest that capturing the meaning of (and thus a fortiori **understanding**) a discourse necessarily involves constructing and using what he terms *discourse representation structures*. He believes that the inter-relation between these partial descriptions of situations is essential for capturing the semantics of tenses, as well as certain uses of quantifiers, including the famous ''donkey'' sentences, such as ''if a farmer owns a donkey, he beats it''.

The use of a representation of the situation under discussion also plays a central role in the more psychologically oriented work of Haviland and Clark (1974), Clark and Haviland (1977), which led them to propose the *Given/New* distinction. They suggest that, even in an entirely

factual discourse, only part of each utterance is intended to directly inform the hearer. The rest is meant to select the entities to which this *new* information applies from those present within an already established, or *given*, context. The syntax of the sentence carries cues for identifying these roles, and its full meaning depends not only on its new information, but also the interaction between its given component and the established context, or *discourse model.* If ever *given* information is found to involve entities that are not already in the hearer's discourse model, that model is deficient. This discovery prompts the hearer to make *bridging* inferences, using background knowledge of the topic to suitably extend the model it until it contains all the mentioned entities.

This notion is greatly elaborated by Sanford and Garrod (1981), who subdivide the information that defines the context, essentially in terms of recency of mention or implication. They then propose constraints on referring to things using determiners and descriptive phrases in terms of the status of entities within these various components. Clark and Haviland's bridging inferences result in the transfer of entities from low to high recency sections of the model when they are referenced as given. However, in sufficiently constrained contexts, some such transfers happen not when the implied entities are referenced, but as part of the processing of the original sentence – in the terminology of Artificial Intelligence, they are *data driven.* Since the transfer process can be expected to take time, Sanford and Garrod carried out experiments intended to make this difference observable. They found that while there is usually a time penalty in processing a reference to implied but unmentioned entities, it disappears when those entities play a sufficiently central role in the situation as already described. Finally, Stenning (1977) has found evidence in errors that subjects make in remembering lengthy sets of directions that strongly suggest that they are basing their recall on a specific, map-like representation of the route.

The next section will review Johnson-Laird's arguments and evidence for identifying these discourse models with the mental models that have formed the focus of the discussion so far. Making such a move means that the restrictions imposed on the basis of mental imagery phenomena can be expected to carry over into discourse interpretation, and Johnson-Laird's experiments will be interpreted as providing evidence of precisely this phenomenon. However, it will also be suggested that while the overall thrust of his arguments is in line with the objectives of constraining the cognitive architecture, there are certain aspects of Johnson-Laird's theorising which leave much to be desired.

3.3. Mental Models and Semantics

Johnson-Laird begins the presentation of his ideas on the processes underlying language comprehension by stressing the importance of capturing the mapping between language and the world.

> The task for psychological semantics is to show how language and the world are related to one another in the human mind – to show how the mental representation of sentences is related to the mental representation of the world.
>
> (Johnson-Laird, 1983. P232)

On this basis, he outlines (Chapter 10) and finds fault with styles of psycholinguistic theorising based on the widely-used notions of semantic markers, semantic networks and meaning postulates.

He concludes that

> Despite their differing inadequacies, the most important characteristic of the three theories is what they have in common. Their main scientific function is to account for the perception of semantic properties such as anomaly and ambiguity, and semantic relation such as synonymy and paraphrase. They are silent on the question of how language is related to the world.
>
> (ibid. P230)

These approaches each rely on capturing the meaning of an utterance in some kind of internal language, and their theorising is focussed on the business of translating into it. Crucially, this process is seen as independent of the actual semantics of the internal language, the specification of which indeed is usually avoided. Indeed, by the choice of suitably "meaningful" names for its elements, this aspect is usually left as an exercise for the reader.

Johnson-Laird terms this the principle of *psychological autonomy of intentions* – "the thesis that semantic properties and relations are grasped by processes that operate independently of those that mediate reference" (Johnson-Laird, 1983. P232). His objection to any kind of theorising based on a language in the head is motivated by his belief that this thesis is untenable. He holds that the extraction of the relations expressed by an utterance often depends upon the properties of the precise individuals named within that utterance – that it is only possible to discover what a sentence is saying by considering what it is about. He supports this with a discussion of a range of experimental results and linguistic observations.

In the normal course of understanding an utterance, a single message is readily extracted from a sentence. To accomplish this, the processing system must usually select the correct reading for each of several words that have more than one meaning. Johnson-Laird suggests that this process depends on the details of what is being talked about. Thus "Michelangelo painted the infant Jesus" and "John painted his bathroom" clearly differ in the sense of the verb that is involved, not least in the final location of the paint. Determining which is appropriate in any particular case of "X painted Y" involves taking account of everything known about both the artistic leanings of X and the surface texture of Y.

However, context can do more than just select among possible uses of a word: there is evidence that it can increase the specificity of the contribution that the chosen one makes to the meaning of the sentence. Johnson-Laird cites the experimental results of Anderson et al. (1976), Anderson and Ortony (1975) and Garnham (1979), which examine the effectiveness with which probe words elicit the recall of sentences. The most obvious way to bring a previously presented sentence to mind, which might appear to be potentially the most effective, is to present a word it contains. However, these experimenters found that it can be more effective to present not a word in the original sentence, but one expressing a specialisation of such a word. For instance, Garnham tested the pair of sentences "The housewife cooked the chips" and "The housewife cooked the peas". He found that "fried" was a better cue than "cooked" for the first sentence, but the opposite was the case for the second. This suggests that whatever representation of the sentences that people commit to memory contains more information than the individual words in the sentence

conveyed.[1] Crucially, this specialisation involves the details of what the sentence is about – recognising that chips are fried but peas boiled is not a matter of linguistics, but of making use of empirical knowledge of cooking vegetables.

Similar effects can also be obtained with verbs, as Johnson-Laird points out by citing the work of Nunberg (1978) on sentences like "The ham sandwich is getting impatient". Clearly, "getting impatient" is not one of the things that things described as "the ham sandwich" normally do. Nevertheless, such an utterance would make perfect sense in the context of a waiter bustling into the kitchen of cafeteria, where it would clearly mean that the person who ordered the ham sandwich was growing restless. Moreover, while the meanings of the words of the sentence determine that the literal interpretation is inappropriate, it is the context that indicates how it should be "repaired". It is not a case that there is simply a standard, syntax-driven "repair" procedure that can generate the canonical non-literal interpretation. The same words could doubtless be used to suggest that the last ham sandwich in a shop was going stale, or to chivvy a reluctant child to finish his tea, or to hasten a dithering diner to take his meal away from the cash-desk to eat.

Finally, over and above Johnson-Laird's examples of subject matter selecting or refining a meaning, it can also be used to **create** one "on demand". Dungeons and Dragons is a game in which players command the actions of characters who inhabit a world where fearsome and magical monsters must be fought with both swords and sorcery. Within this mythical world, as indeed in many universities, there are those who, by diligent study, have learned how the appropriate recitation of certain incantations can act as an irresistible soporific. When players discuss the use of these awesome spells of sleep, they spontaneously speak in terms of Dan the wizard trying to *sleep* the largest Gnome. Despite the fact that "sleep" is usually held up as an archetypal intransitive verb, it is spontaneously used transitively if circumstances suggest that this is appropriate. Indeed, this process must have been involved with the origins of phrases like "tabling" a motion, "chairing" a meeting or "booking" an offender.

These arguments lead Johnson-Laird to reject the notion that human language processing involves a stage of syntactic transformation into an intermediate formalism. Instead, he suggests (Chapter 13) that these phenomena indicate a mechanism that directly translates from (something very near) the surface form of English directly to a semantic representation – namely, a mental model. Only by doing so does it have the information necessary to accommodate the findings presented above. It is this use that other, non-linguistic, modules make of the mental model that gives it its semantics. It is what distinguishes it from the uninterpreted symbols of logicians, and what Johnson-Laird believes distinguishes human mental models from the data structures of a computer program:

> Human beings know how to relate expressions to models of the world; unlike the program, however, they have ways of constructing such models that are not dependent on linguistic input.
>
> (ibid, P263)

The fact that other modules mediate its interaction with the world is what allows the construction

[1] This result has an obvious similarity to Sanford and Garrod's finding that some bridging inferences are data-driven.

of a mental model from a sentence to be called understanding.

Moreover, characterising discourse models as mental models that are accessible to other (e.g. input) modules also allows a natural explanation of the processes underlying the phenomena of *deixis*, the use of words to refer to the immediate environment. Pronouns, such as "they", are most often characterised as words that refer to (groups of) entities previously mentioned (or – the whole point of Haviland and Clark – implicated) in the discourse. However, they can also be used without connection to any prior use of language. Reichgelt (1986) illustrates this well-known phenomenon very neatly with the example that, given only the context of passing a particularly luxurious house, it is permissible to start a discourse by wondering "how much do **they** earn?". The referent of "they" is clearly the owners of the house, who have previously been neither mentioned nor implied, and indeed have not even been seen. The use of the pronoun can only be justified by the expectation of bridging inferences based on the perceived situation and the fact that houses usually have owners and large ones are expensive to run.

Johnson-Laird considers that one of the vital features of this approach is that

> The processes by which fictitious discourse is understood are not essentially different from those that occur with true assertions.
>
> (Johnson-Laird, 1983. P246)

This notion was already touched on in the discussion of the treatment of imaginary and perceived situations, from which the extent of the consequences of rejecting it should be clear. However, accepting it is unproblematical, since it is equivalent to a denial of omniscience. The hearer, and indeed the **speaker**, of any particular utterance has to select and execute an appropriate processing strategy without any way of knowing whether it is true or fictitious.

The psychological autonomy of intentions is not the only feature of standard psycholinguistics that Johnson-Laird opposes. He also rejects the notion that linguistically expressed concepts map to atomic forms in some internal language with arbitrarily convenient semantics. Instead, his commitment to the perceptual connections of mental models leads him to propose that entities are represented as complex, multi-featured, descriptions, which he likens to the notion of a *frame* as discussed in (Minsky, 1975). Johnson-Laird explicitly traces this idea back to the Gestalt tradition, citing Smoke

> As one learns more and more about dogs, his concept of "dog" becomes increasingly rich, not a closer approximation to some bare "element"... No learner of "dog" ever found a "common element" running through the stimulus patterns through which he learned.
>
> (Smoke, 1932. P5)

This argument is, of course, mirrored in Wittgenstein's famous discussion (Wittgenstein, 1953. P31-32) of the "family resemblance" that holds between instances of games, and in Minsky's (1967) characterisation of "machines". It also finds empirical support in the experimental work of Rosch (1976), who proposed that natural categories are represented within the mind as *prototypes*, which serves to emphasise that the values of frame slots act not as necessary characteristics, but as default values.

Within a prototype-oriented system, the applicability of a category term to a particular entity would not be a matter of a precise (but arbitrary) predication. Instead, it would be a quantitative

measure of similarity between the entity and the relevant prototype. As such, and unlike the boolean certainties of traditional logics, it is a system of categorisation that at least has a chance of saying anything at all about the real world. The answer to "Is France hexagonal?" is neither "yes" or "no", but "fairly"! Moreover, the relative significance attached to the differences in each specific feature could be varied in line with the purpose of categorisation. The system at least takes account of the right kind of factors to account for so many logicians' burning desire to call a packing case a table one moment and a chair the next.

To this, Johnson-Laird adds the notion, borrowed from de Saussure, that

> The boundaries of a word's extension are set, not by its schema, but by the nature of the particular taxonomy in which the schema occurs. Whether something is to count as a dog depends on its similarity to typical dogs, typical cats, typical foxes, and so on.
>
> (Johnson-Laird, 1983. P202)

Johnson-Laird points out the value of such a notion in a situation where new concepts are being developed or, of fundamental importance for language acquisition, being learnt. It is an approach that functions well even when, as he puts it, "the 'rules of the game' are incomplete", and naturally allows the use of an ill-fitting term, such as "Iron Horse", if there is nothing more suitable available.

In essence, Johnson-Laird's proposed language processing mechanism interprets a sentence by modifying an existing discourse (mental) model in order to make that sentence true. Often the modification will require adding new properties to (or relations between) entities already present. Alternatively it may involve adding entirely new entities, as is sure to be the case at the start of a discourse, when the existing mental model will be completely empty. In any event it will be semantically guided, in that the procedures that make it will be working with a representation geared towards encoding all available information about the subject matter. From the viewpoint of predictive theories, particularly with a view to being able to embody them in programs on currently available computers, this much flexibility is a serious deficiency. So too is the notion of uncertain assignment to ill-defined categories. However, Johnson-Laird's arguments are aimed to establish that they are also essential features of any adequate account of human linguistic behaviour.

One of the most powerful features of the proposal is that it allows many bridging inferences to be made automatically. Any model constructed in line with a description will necessarily fit the applicability conditions of any other sentence it (normally) implies. Thus in the normal interpretation of a description like "a woman standing at a bus stop", the information structure added to the blackboard in response to the phrase "a woman" will fall within the criteria that determine the applicability of the term "person". Equally, it will definitely fit the criteria for "living creature" and "physical object", grossly mis-match with those for "dog" and "sitting" and partly match those for "immobile" and "upright". More importantly, it will also be amenable to evaluation in terms like "heavier than a hamster", despite the fact that such a comparison is unlikely to arise outside a psychology laboratory. It will also contain information about sub-structures that can be described as "arms", "legs" or "head", warranting referencing those parts of her anatomy as given information.

Similar phenomena can be observed relating to other kinds of immediate inference. Johnson-Laird focuses on the particular example of those concerned with the relative positions of objects, such as the layout of cutlery on a table. He points out that when people are told that "the spoon is to the right of the knife" and "the knife is to the right of the fork", they immediately recognise that this implies that "the spoon is to the right of the fork". He suggests that this is not the conclusion of any explicit inference involving the transitivity of "to the right of" – that is, no representation of the inferences that this property warrants plays any part in the solution process. Instead, it is the result of the fact that the initial interpretation of the sentences involved building a model of the situation described, which required arbitrarily assigning appropriate positions to the objects mentioned. These positions were represented in a manner that ensured conformity with the properties of spatial relations. This could either be because of the physical properties of the representation involved (i.e. the use of a truly analogue representation) or because of the manipulation of them was strictly controlled. Of course, the choice of such a representation demonstrates (implicit) knowledge of the relevant properties. Nevertheless, no explicit (cognitive) inference is required to recognise the conclusions – they emerge from the representation of the situation.

Similar evidence against the involvement of explicit inference in the appreciation of transitivity comes from the a kind of Symbolic Distance Effect that can be observed in transitive inference tasks. The term was coined to describe a phenomenon concerning the retrieval of certain kinds of information from long-term (common sense) knowledge. Moyer (1973) and Paivio (1975) report that the time subjects take to judge the relative size of two named[2] objects decreases with the size difference between them. Thus they found that subjects will judge a cat to be smaller than an elephant faster than they will compare it to a dog. Interestingly, a very similar reaction profile is observed in transitive inference tasks, such as that carried out on infants by Bryant and Trabasso (1971) in their attempt to test Piaget's suggestion that they are incapable of tackling such complex problems. The experiment was subsequently replicated, using monkeys as subjects, by McGonigle and Chalmers (1977), who obtained essentially similar results.

In each case, subjects were initially trained (rewarded) until they could reliably demonstrate knowledge of an arbitrary (that is, with no perceptible manifestation) relation between each of the the immediately neighbouring pairs in a series of 5 items. They were then presented with non-adjacent pairs from the series, which they had not been trained on, and immediately made the appropriate choices – that is, having been rewarded for choosing green in preference to blue, and blue in preference to red, they spontaneously chose green in preference to red even though they had never been trained on that pair.[3] Crucially, subjects responded faster to non-adjacent pairs on which they had not been trained than they did to the adjacent ones on which they had. This is the opposite of what would be expected if the results of the training were being remembered as a collection of pairwise relationships and responses were being made on the basis of any kind of

[2] The effect remains if the subjects are shown uniformly sized (i.e. not to scale) pictures of the objects.

[3] It is necessary to use a five-term series to ensure that it is possible to present a pair that they have not been trained on, each member of which has been both the preferred and rejected member of a training pair.

explicit use of transitivity. Such an approach would require at least one additional and presumably time consuming inference step for each intervening item. Instead, it seems reasonable that as subjects – either infants or monkeys – are being trained, they are not simply noting the individual pairings, but are constructing a representation in which the complete ordering is directly available if required.

Moreover, the encoding of the interactions between objects that allows inference to be avoided can be shaped, like that between word senses, by knowledge of the meanings of the words involved. To see this, consider what is conveyed by being told that "the clock is in the centre of the mantlepiece, and John's photograph is just to its right, carefully positioned to hide a nasty mark on the wall". It is, of course, immediately obvious that the mark is on the wall behind the clock, just to its right. However, this inference, and indeed the use of a definite article to name the wall, can only be justified on the basis of knowledge about mantlepieces and where you have to put a picture to hide something. In addition, as a mental model it is available for manipulation and extension by all processing modules. This means that there is a smooth transition between the kinds of static implications discussed so far and the kind of dynamic consequences that are no less immediate, such as that required for understanding the definite reference in the following

> As they tiptoed past the slumbering gunmen, his companion's scarf snagged the crystal decanter, dragging it from the mantlepiece. His heart in his mouth, Bond's hand flashed out and grasped **the falling glassware**.

This makes clear that the model, being the place where the system's specialised knowledge (which includes its "common sense") is brought to bear, is sharply distinct from the state or output of any individual module. Fodor (1983, PP 43-46) emphasises the difference between the offerings of the input modules and the description of the world relevant to the fixation of beliefs. He stresses (P89) the fact that the latter are the "best" explanation of the state of the world that the cognitive system can come up with "all things considered", and "all things" is precisely what an informationally encapsulated input module is unable to consider.

Of course, this doesn't mean that a discourse model necessarily reflects the contribution of every available piece of knowledge about the world. The evaluation of its interaction with stored information would be carried out by processing modules, and as a result, its extent would be limited. Obviously, the information content and processing capacity of the modules imposes restrictions, but it is important to realise that since these modules are domain specific, their apparent relevance also has implications. Processing modules can only apply their knowledge to a situation that is presented in the terms that they can "recognise". Thus when people interpret a discourse on familiar or everyday matters, such as the arrangement of cutlery on a table, they bring to it a sophisticated processing system customised to manipulate just such information. It will be able to ensure that any model built is consistent with its knowledge of how the world is and works. On the other hand, few people can bring such rich experience to a discussion of the more complex failure modes of a computer operating system. In such situations, the language system is essentially "on its own". It still creates an information structure that combines the meanings of the words, but there is no processing module that can check its plausibility or flesh out details or implications. Even though the individual words may be known – possibly even in the sense of

having a definition associated with them – the discourse is at best only partially understood.

Unsurprisingly, there are strategies that can be adopted to ameliorate the problems of having to deal with barely comprehensible situations. One possibility is to express the situation in more familiar terms that will engage suitable support processes – in short, to use *metaphor*. As a result, the ubiquitous and crucial nature that Lakoff and Johnson (1980) identify for metaphor should be no surprise. In the case of the innards of a computer, it is usual to think of data structures situated at *addresses* in some kind of *space* – indeed, it is even commonplace to draw a memory *map*. These familiar spatial notions bring with them ideas of separation, extent and capacity, not to mention the notion of being able to find at most a single thing at each point. Within this world exist entities – processes – which are spoken of as having purposes and knowledge, which leads them to modify and rearrange data objects. Once again, actions and desires, although they may defy the logicians attempts at formalisation, are things with which everybody is familiar since infancy. It is no accident that the most-used feature of the most used programming language is called a "move" instruction. It is part of an elaborate metaphor that has a vital role in understanding, and thus working with, the operation of these phenomenally complex machines. It can play it because it brings the problem within reach of processing modules.

3.4. Discourse Coherence

These arguments suggest that the use of a mental model will significantly improve the way a discourse is handled, and that anything that disrupts its smooth construction will impair performance. In section 2.4 it was suggested, on the basis of the interference between perception and imagery, that the mechanisms underlying mental models have severely limited capacity to deal with more than one model at a time. When these notions are combined, it leads to the prediction that a discourse will be harder to process if the sentences within it do not integrate smoothly into a single model. In particular, if a sentence refers to an object already introduced into the model by a previous sentence, this significantly constrains how the rest of the information it contains should be interpreted. On the other hand, if it refers to nothing that has been mentioned (or implied) before, it cannot be smoothly integrated and must in some sense be kept separate.

Ehrlich and Johnson-Laird (1982) report an experiment that demonstrates precisely this effect in the interpretation of descriptions of the layout of four items. These contained three sentences, each specifying the relationship between two objects (e.g. "the cup is on the left of the plate"). Since such relations can be expressed with the referring terms in either order (by reversing the relation between them), there are 8 different triplets of sentences that all present the same information. Within each triplet, the sentences can be ordered to produce a description that is either referentially continuous (the second sentence referring to something in the first, and the third referring to something in the second) or discontinuous (i.e. the first two sentences being unrelated, the last linking them by referring to something in each). Thus, for instance,

(i) The knife is in front of the pot.
(ii) The pot is to the left of the glass.
(iii) The glass is behind the dish.

constitutes a continuous description, whereas if the sentences were presented in the order (i), (iii),

(ii) it would be discontinuous, since sentence (iii) in this position would not refer to anything in the established discourse – i.e. sentence (i).

This combination of eight distinct triples ordered to give both states of continuity gives a total of 16 forms, and the experiment involved one example of each being read to each subject. Subjects were able to indicate when they were ready to hear the next statement, and at the end of the triplet had to draw a diagram of the layout that had been described. Ehrlich and Johnson-Laird found that continuous descriptions were handled much more easily (57% vs. 33% accurate) and faster than discontinuous (mean listening times 13.3 vs. 14.9). Most errors involved forgetting an item, with a total of 40 items being omitted from diagrams made from discontinuous descriptions, as opposed to only 13 from those made from continuous descriptions.

While this result strongly indicates the importance of being able to construct a single representation of the entire discourse, Ehrlich and Johnson-Laird attempted to probe its internal organisation. In particular, they opposed the notion, put forward by (Kintsch and van Dijk, 1978), that the relevant information was represented as a list of propositions, ordered to reflect their introduction to the discourse. One of the fundamental properties of mental models that is suggested by mental imagery is the uniform accessibility of its entire contents. They therefore extended the experiment presented above in order to investigate the effect of varying the delay between the reference to an object and its introduction to the model. Specifically, they introduced a third ordering of the descriptions, in which the third sentence did not refer to anything in the second (as it would in a continuous description), but both referred to an entity in the first. This would be the case if the example sentences given above were presented with sentence (ii) first. This third form gave them a total of 24 test items, which were again read to subjects, this time at fixed four-second intervals. As before, discontinuous descriptions produced considerably less accurate diagrams than those from continuous descriptions (42% vs. 69% accurate), while semicontinuous descriptions did not significantly degrade performance (61%). This suggests that the availability of explicitly named features in the model is not dependent on their recency of introduction.

Finally, Ehrlich and Johnson-Laird probe the relative accessibility of features within a model not in terms of processing errors, but directly in terms of the durations of the processes involved. As before, they presented subjects with the eight combinations of sentence forms in each of the three states of referential continuity. However, in this experiment the sentences were not read aloud but presented visually, one at a time, using a computer terminal. Subjects indicated their readiness to read the next sentence by pressing a key, with the time spent reading each being noted by the computer, and at the end of the third sentence they drew a diagram of the object layout. As before, continuous and semi-continuous descriptions produced similar accuracies (63% and 61%) which were significantly better than discontinuous (40%). A similar pattern emerged in the average reading times per sentence, where they were again comparable (6.1 and 6.3 seconds) and significantly faster than those for discontinuous descriptions (7.1), a difference which arises almost entirely because the last (linking) sentence took an average of 9.4 seconds.

Ehrlich and Johnson-Laird summarise these results by saying

> The trend over the discontinuous sentences suggests that the second sentence slows subjects down because if fails to refer to any previous object, and that the third sentence takes still longer because not only must it be interpreted but it must also be used to integrate the information conveyed by the two previous sentences – only at this point can the subjects construct a unitary representation. The results cannot be explained by assuming that subjects form co-referential links between ordered propositions, because the cyclical process of searching through the buffer would yield a systematic difference in the times taken to interpret continuous and semicontinuous descriptions.
>
> Ehrlich and Johnson-Laird (1982, P303)

This summary is a rather problematical. The published figures show the average time to interpret the (referring) second sentence in a continuous description is more than one standard deviation **slower** than the corresponding time for a discontinuous description. This makes it hard to justify their suggestion that subjects slow down when confronted with a non-referring sentence. Moreover, the prediction of an effect of buffer searching is trivially dismissed by suggesting that the working memory buffer could well be so small as to be searchable without observable delay, or could support parallel-access or even content addressability.

Despite this, the central claim concerning the importance of referential continuity appears sound. It is clearly very beneficial if each sentence links to some item already established in the discourse, and that this benefit does not depend on the immediacy of that entity's introduction. Moreover, attributing a central role to a mental model provides a natural explanation of why this should be so. Discontinuous descriptions are hard to deal with because they cannot be handled by the smooth enhancement of a single model of the situation. Since juggling two models is extremely difficult, subjects can only handle them by adopting some other, unrelated and less effective mechanism. Specifically, they postpone the processing of the problem sentences, and merely attempt to retain it in some uninterpreted form in order to have it available for integration when a suitable linking sentence is encountered. This shows itself in the additional processing observed to follow the reading of the final (linking) sentence of the discontinuous descriptions.

Because of the difficulty of handling discontinuous sentences, it is reasonable to expect that normal discourse will be structured in order to avoid them. Indeed, Johnson-Laird goes so far as to make their absence a **criterial** for a well-formed discourse: He proposes that

> A necessary and sufficient condition for discourse to be *coherent*, as opposed to a random sequence of sentences, is that it is possible to construct a single mental model from it [sic] ... Each sentence must refer, explicitly or implicitly, to an entity referred to (or introduced) in another sentence, since only this condition makes it possible to represent the sentences in a single integrated model.
>
> (Johnson-Laird, 1983. P370)

In many ways this proposal is simply a restatement of the obvious fact that a sensible discourse must be **about** something. Its sentences must contribute information about the same objects, which means that they must refer to things that have already been mentioned.

Given that they are linguistic tools specialised for just this purpose, the way pronouns are used is clearly important to characterising a well-formed discourse. However, it is not enough, since definite noun phrases also refer to something that has already been mentioned, so they too can hold a discourse together. Crucially, determining whether two noun phrases can co-refer requires taking account of their full semantics which, Johnson-Laird argues, means using the discourse model. This is illustrated by the discourse

I caught a kid climbing on my car today.
I told the X that if it happened again I would clobber him.

If X is a phrase like "little tyke" or "brat" then the second sentence is clearly elaborating the speaker's interaction with the offending child at the time of the offence. However, if it is "policeman" or "man on the ground floor" it is probably discussing a completely unrelated episode – namely, a subsequent conversation with another individual, and thus signalling that a new discourse model may well be appropriate. Which interpretation is correct can only be evaluated on the basis of precisely what X refers to.

It appears, then, that considering the structure of a discourse leads back to Johnson-Laird's central claim that mental models are vital to language comprehension because they are an information structure which is semantically interpreted by the cognitive system. As such its role in determining noun-phrase referents not only enriches the interpretation of the specific sentence, as described above, but controls the way the discourse holds together. Indeed, the search for a suitable co-referent serves both to signal that bridging inferences are necessary and that they have been successful. As a result, the requirement that coherent discourse should be interpreted by the enhancement of a single model gives rise to both the its overall structure and to surface features of the kind that inspired Haviland and Clark and Sanford and Garrod.

While definite noun phrases essentially call for a search for an already mentioned or implicated entity that fits the relevant description, indefinite descriptions indicate that this search will be fruitless, and a new entity should be introduced. The model should be extended by constructing, from the prototypes associated with the elements of the indefinite description, the representation of an entity typical of the things that it can describe. Thus the mention of "a big dog" is a first reference to a canine about which the speaker assumes that the hearer knows nothing at all. As such, it commands the hearer to extend his understanding of the situation under discussion by including within it a (or another) representation of a typical "big dog".

Unfortunately, striving to apply this interpretation uniformly to all indefinite descriptions seems to have led Johnson-Laird to propose grossly distorted interpretation of predicative sentences. This is illustrated by his proposed treatment (Johnson-Laird, 1983. P383) of the sentence "Ann is a teacher". He suggests that interpreting it involves constructing a mental model containing not only an entity description for Ann, but also for an arbitrary number of typical teachers, one of which is to be shown in an "identity relation" to her. He then suggests that it is the presence of the other teachers – those not involved in an identity relation with Ann – that prevents this sentence being followed in a discourse by "the teacher caught a bus".

This argument is unconvincing, if only because it is very easy to create a context in which "a teacher" can subsequently be referred to by a definite article:

There are two men.
One is a teacher, the other is a miner.
The teacher...

Moreover, it also suffers from the even greater unacceptability of continuations which would be appropriate for a model containing a number of teachers. "Ann is a teacher" cannot be followed with either "One of the teachers caught a bus" or even "Some of them caught a bus". The

anomalousness of Johnson-Laird's discourse is much better explained in terms of the Gricean principle of cooperation, which is being flouted by the failure to use the more direct ways of referring to her – i.e. her name or a pronoun. Indeed, it is arguable that in these circumstances, particularly since there are no possible conflicting interpretations, the convention of pronoun use is almost obligatory (cf. John is cleaning John's teeth).

The real root of Johnson-Laird's problem is in the attempt to introduce new entities. What is actually required by a predication is the modification or enhancement of an established representation to make it a description of a typical, or at least passable, possessor of that property. The true meaning of a statement such as "Ann is a teacher" is to convey the applicability to the entity named by "Ann" of the information normally relevant to members of the category "teacher". Understanding it requires that those properties be appropriately integrated within the information structure describing Ann, and the problems of any resulting conflicts resolved, or at least discovered. For instance, anybody who held that the typical woman was a housewife whose life consisted of shopping, cooking and a game of Bingo on Friday night would be obliged to revise the picture of Ann that they would otherwise have assumed. This process is vital to the ability of mental models that allows them to support the kind of semantic effects that Johnson-Laird sees as so important. In replacing it with the mere insertion of an "identity link", he is undermining the whole enterprise and making mental models just another form of semantic net.

The problem is even more acute in Johnson-Laird's handling of universal attributions. He proposes (Johnson-Laird, 1983. P391) that the interpretation of a sentence like "every man loves a woman who loves him" involves the creation of a model containing an arbitrary number of prototypical men suitably enamoured with equally adoring women. Such an interpretation of the sentence may capture precisely the kind of meaning that logicians profess to believe it conveys, but it is totally unrelated to its contribution to any conceivable discourse. Its meaning is much better captured by recognising it as information concerning the properties of the prototypical member of the category. Moreover, since many properties of such prototypes are only "defaults", this interpretation can naturally captures the essence of many of the more elusive uses of such statements. Even in the face of a notice-board full of counter-examples, it is still possible to meaningfully assert that "all students plough their first year examinations", since its true message is that the hearer should note that a typical student is quite likely to do badly. The same approach is clearly able to take account of the plethora of weaker quantifiers that can be used in such attributions (most, hardly any, absolutely all). These matters are discussed at length, with particular emphasis on the so-called "donkey" sentences, in Reichgelt (1986), using the notion of an arbitrary object proposed by Fine (1983).

3.5. Vagueness

There is a potential difficulty with the suggestion that indefinite descriptions trigger the addition to the mental model of the representation of an arbitrary entity with the appropriate properties. The Section 2.4 proposed that their close connection with imagery made it natural to characterise mental models as having particular restrictions on the things that could be represented with uncertainty. This is a problem, since merely calling for the representation of an object with a

certainly property is laden with ambiguity – what colour is "a big dog"? Indeed, it is probably the rule, rather than the exception, that language conveys only incomplete information, even about the very aspects it is explicitly dealing with. For instance, the sentence "Mr. Smith lives nearer the shops than Mr. Brown, and so does Mr. Baker" leaves unclear the relative lengths of shopping expeditions for Messrs Brown and Baker. This suggests that the identification of the kinds of models used to capture discourse with the information structures underlying mental imagery should have observable effects concerned with the interpretation of ambiguous discourse.

Mani and Johnson-Laird (1982) report two experiments that show precisely this effect. They have clear similarities with those of Ehrlich and Johnson-Laird (presented above), being based on presenting subjects with a sequence of sentences that each specified the spatial relationship between two of a set of items. However, instead of having to draw the arrangement of objects at the end of each description, subjects merely had to indicate whether a particular drawing they were shown was consistent with the description they had just heard. Furthermore, the descriptions were all at least semi-continuous, but varied instead in whether they described a unique arrangement of objects. Finally, after the last of the sixteen descriptions subjects were given an unexpected recognition test, in the form of being presented with four sets of sentences, which they had to rank in order of resemblance to the original description. The sets they were given each contained a sentence they had been given in the main test, a paraphrase of it and two other sets of sentences that described a different layout of objects.

Mani and Johnson-Laird found that subjects classified 7.3% of diagrams wrongly, which was unaffected by either determinacy of description or consistency of diagram. The description ranking, however, showed a strong effect of determinacy, with subjects ranking both the original description and its paraphrase above the two distractors significantly more frequently for determinate descriptions. In other words, the overall situation was remembered better when the description was determinate. The origin of the effect is suggested by examining subjects' preferences between the original description and its paraphrase. Although there is no effect overall, there is a significant difference when the analysis is restricted to items where the subject remembered the content – that is, where both were ranked above the confusion items. In these cases, the original description was ranked above the paraphrase significantly more often than chance, but only for indeterminate descriptions.

This result is very difficult to explain without postulating the evaluation of the meanings of the premises of the kind that building a model involves, since the determinacy of a description cannot be determined solely from the syntactic features of its propositions. It is also clearly related to those of Ehrlich and Johnson-Laird, who also found that an inability to continue the construction of a single model results in a change of processing. In their case the problem was the restriction to only a single model, and the result was the adoption of processing based on the manipulation (storage) of an uninterpreted form of the premises. In the current case the relevant restriction is on the representation of the appropriate ambiguity, and a similar increase in processing of the individual premises thus seems a natural prediction. Were it adopted, it would be expected to enhance the ability to distinguish an indeterminate description from a paraphrase, which is precisely

what Mani and Johnson-Laird observed.

A key feature of this experiment is that the recognition test was unexpected, and thus subjects could not have taken any special measures to improve their ability to recall the individual problems. It thus seems reasonable to assume that their recollection of problems must be based on the processing they carried out to solve it. If this is the case, the suggestion that an indeterminate description disrupts model-based processing implies that the proposed explanation of the determinacy of a problem qualitatively affects the information structures underlying subjects' recollections of it. In the case of an indeterminate description, there will have been additional processing of the individual premises, which enhances the ability to recognise their precise wording. This will not have happened with determinate descriptions, where processing will have focussed solely on the construction of a suitable model, which will therefore play a comparatively greater role in description recognition.

Since it simply represents a state of the world, a model is equally compatible with any description of that situation, and in particular would not help to distinguish the the description that led to its construction from any informationally equivalent paraphrase. However, it would be equally unable to discriminate between that description and some other description that contained **different** information about the same situation. This arises directly from the key feature of mental models. They not only represent the information put into them, but also the results of significant processing carried out to enhance it, which in the case of a discourse model would take the semantics of the description into account. This leads to the prediction that subjects should also have trouble distinguishing a determinate description from another that explicitly stated quite different relations.

Mani and Johnson-Laird's second experiment serves to test just this. Subjects were given eight descriptions of the relative positions of five objects which were either determinate or indeterminate. As in the first experiment, subjects were told to classify diagrams presented immediately after the sentences, which they did correctly for 89% of the diagrams, independently of either description determinacy or diagram consistency. They too were given an unexpected recognition test which in this case contained not a **paraphrase** of an original description, but a description that contained different information but was nonetheless **inferable** from it. As before, subjects ranked the original and related descriptions above the confusion items significantly more often when the description was determinate. In addition, however, an indeterminate description once again produced a significant increase in the subject's tendency to rank the original form above the inferable description.

Mani and Johnson-Laird interpret these results as confirming the use of two distinct representations of the discourse.

> The point to be emphasised, however, is that the results from the two sorts of description cross over, depending on whether one is measuring the gist or verbatim recall. This finding cannot be accommodated by any explanation that resorts to only a single unitary notion, such as degree of organisation, amount or information, or abstractness of representation.
>
> (Mani and Johnson-Laird, 1982)

Moreover, they believe that they also reveal something about their contents.

> Granted the need to postulate at least two sorts of representation, it does not follow immediately that one must be a spatial model and the other a propositional representation close to the linguistic form of the description. However, the results of the present experiments thrust such a conclusion upon us... First, the recall of verbatim detail demands access to linguistic form. Second, confusion of the original with one that can only be inferred from a layout constructed in a symmetrical way from the original demands access to a spatial model.
>
> (Mani and Johnson-Laird, 1982. P185 - 186)

Hagert (1985) makes a garbled attempt at presenting an experiment which contradicts these results. Subjects were presented with descriptions of scenes, and "afterwards" asked to either draw the corresponding layout or rank sentences as either seen or new. It is not clear from his report whether "afterwards" refers to the end of the experiment or immediately after the presentation of each item. If the former, there is no indication what task, if any, the subjects were given to do at the time they could see the description. However, the latter seems most likely, and hence subjects must have known (or very quickly realised) that they would be asked about the wording of the descriptions.

Hagert reports that on average 50% of the diagrams that subjects drew were correct, which was unaffected by the determinacy of the description. This is in startling contrast to Mani and Johnson-Laird, whose subjects could classify 93% of diagrams correctly, while on precisely the same task Ehrlich and Johnson-Laird found 70% of drawings correct from subjects who only **heard** the descriptions.[4] Hagert also found no effect on reading times, with his subjects taking an average of 105 seconds to read 4 sentences. This is approximately 25 seconds per sentence, which again is significantly different from the 5-10 seconds per sentence that Ehrlich & Johnson-Laird's subjects required for a similar task, albeit one sentence shorter. Finally his subjects correctly classified between 80% and 87% of the descriptions, and while this is higher than 50%-75% (depending on the original description), it is nevertheless surprising considering that they performed their recognition test at the end of the entire experiment.

Overall it is difficult to assess Hagert's experiment. It appears likely that his subjects knew there would be a sentence recall test while they were studying the premises, and thus adapted their behaviour to concentrate on this aspect of the task. If so, it is not surprising that their performance was unaffected by the structure of the contents of the descriptions. The objective of Johnson-Laird's experiments is not to discover whether subjects can create a representation that is detailed enough to support any kind of interrogation. Instead, it aims to probe the properties of the representation that they will use if they think (wrongly) that they can optimise for one particular task.

In addition to reporting experiments, Hagert (1984, 1985) is also responsible for one of the most direct assaults on Johnson-Laird's idea of mental models. Hagert (1984) reviews the experiments that Johnson-Laird uses to support the idea of mental models (Ehrlich and Johnson-Laird, 1982, and Mani and Johnson-Laird, 1982), but takes issue with the claim that they show that the mind employs distinct representations. He argues that the effects of continuity of reference and

---

[4] Admittedly this fell to 33% for discontinuous descriptions, but it is not obvious that this will be relevant to Hagert's subjects, who could **read** the premises.

determinacy can be achieved using what he believes to be a single, uniform, representational system.

To demonstrate this, Hagert outlines a form of representation based on propositional networks, with working memory alleged to correspond to those propositions that are "activated". A central feature of this notation is the idea of a *composite proposition*, such as

Left(front(Knife, Pot), front(Dish, Glass))

which is intended to capture the layout of objects described above. However, as Anderson (1978) is at pains to point out, no representation is complete without the procedures that manipulate it. Apparently inspired by the production system OPS (Forgy and McDermott, 1979), Hagert describes these in the form of rules, though without clarifying whether these specify re-write (i.e. replacement) or inference (i.e. incremental) operations. These rules, he points out, "can form larger sources of knowledge, e.g. schemas realising the meaning of a particular concept" (Hagert, 1984. P393). He illustrates this by presenting ten rules which encapsulate the concept of "left", some of which are illustrated in Fig. 3.1. Of these, three serve to construct *composite propositions* (e.g. R1), another three to break them up again (e.g. R5) and three more to reflect the interrelation between left and right (e.g. R9). The last states that "front", which is mentioned even though the rules are for the concept of "left", contradicts "behind". Crucially, the process of combining and dismantling these composite propositions can give rise to new information that was not given in the original premises. Thus if Left(x, y) and Left(y, z) are combined to give Left(x, left(y, z)), (by R1) this can then be broken back down to yield both these original premises **and** a new proposition Left(x, z)

Having outlined these ideas, Hagert then goes on to show how these rules, when manipulating composite propositions, can predict the effects that Johnson-Laird observes. The detrimental effect of discontinuous reference arises because by the time the linking premise is being heard, the representation of the first is just on the point of fading from working memory. The effect of determinacy of description, both on sentence recognition and content recall, is more complex. The rules of the schema can (and will) build up a determinate description into a single composite proposition. However, this process cannot be completed for indeterminate descriptions, and this, together with the fact that the incomplete process is shorter, explains the observed effects. Why this should be so is not clear, and unfortunately Hagert's only explanation is a cryptic allusion to the fact that things fade from working memory. Diagram evaluation is better for determinate descriptions because the single composite propositions that they produce are better able to match

---

R1: Left(x, y) & Left(y, z) -> Left(x, left(y, z))
R5: Left(front(x, y), z) -> Left(x, z) & Left(y, z) & Front(x, y)
R9: Right(x, y) -> ~Left(x, y)

Fig 3.1: Some of Hagert's rules for the concept of "Left".

---

against the diagrams presented. However, the explanation is vague and seems crucially dependent on the details of the rules in the schema.

Hagert's theory is beset with mundane problems and curious assumptions. For instance, he supposes that inactive elements fade from working memory after a number of rule cycles, rather than after a set period. This may seem natural within a production system mechanism, where a "cycle" is the basic unit of activity. However, there is nothing to indicate that the disparate kinds of activity that occur during one of Hagert's cycles – e.g. reading and understanding an eight word sentence – should be seen as taking the same time as the fastest inference. Moreover, it is hard to see how this can bear the weight of explaining the effect of determinacy of description on verbatim recall of the premises. Johnson-Laird's recall tests were done at the end of the entire batch of test items, and "working memory" would have to have been re-used many times over for the intervening items.

Hagert (1985) goes beyond merely offering an account of a particular phenomenon, and attempts the laudable task of setting this idea in the context of a more general picture of reasoning (see the discussion of (Newell 1981) in Chapter 5 below). He analyses the control of mental processing into three levels. The *domain level* carries out the actual steps in the task under the direction of the *strategy level*, which is designed to suit the task by the *planning level*. Thus he describes the solution of a problem as resulting from a planning level design process, which builds a *conceptual model* (not to be confused with Johnson-Laird's conceptual **mental** models) containing the relevant domain knowledge and a strategy of employing it that is appropriate to the task. This model is then "run", much like a computer program, which might either solve the problem or be redesigned by the planning level and run again.

In this context, the representations and schemata of rules compose the domain level. As before, indeterminate descriptions are shown to differ from determinate ones because they cannot be processed into a single composite proposition. In 1984, this difference meant that "the computation is shorter" for indeterminate descriptions and

> many more inferences and composite propositions are obtained when the description is determinate. [This] predicts that subjects should remember the actual sentences far better if the description is indeterminate, whereas they should recognise layouts and inferred sentences more easier [sic] if the task is determinate. This is also the result from the[5] psychological experiments.
>
> (Hagert, 1984. P396)

However, by 1985 this difference has lost its significance. The same comparison of description types still

> shows that both processes are very similar, but that **processing effort is greater when the task is indeterminate.**
>
> (Hagert, 1985. P276)

which now means

> The very same general prediction holds for the task of recognising sentences among a set. This is in contrast to predictions found in, for example, [REFS[5]]
>
> (ibid)

Hagert then nearly reports an experiment that supports this new prediction (see above). As pointed out there, these results differ from those of Johnson-Laird et al. in several ways. Additionally, there is no effect of determinacy on reading time, about which Hagert makes no comment, despite the emphasis he placed (in the above quotation, quite literally!) on the fact that indeterminate descriptions require (and permit) less processing. However, as suggested above, there is serious doubt concerning what Hagert's subjects thought they were doing, and this could well have had an effect on their reasoning and representation strategies. As a result this experiment, and the theoretical "predictions" made to explain it, are best left to one side.

Finally, the real problem with Hagert's attack on systems based on mental models, which is sufficient to render it totally futile, is that his counter-proposal simply is such a system. He proposes that information on certain aspects of a discourse (e.g. relative positions) will be fully manipulated to extract the consequences of its interaction with the relevant body of stored knowledge. This is, of course, precisely the characterisation of a mental model. Moreover, within such a system, the information in any particular statement – for the sake of logicians, any particular proposition – can either be committed to the model, in which case its consequences will be derived, or held apart and "uninterpreted". Thus although both sorts of knowledge can be expressed in the same formalism – which is unsurprising, since literally anything can be written as some kind of "logic" – two clearly distinct behaviours remain possible. Which is, of course, precisely what Johnson-Laird has argued for and sought to demonstrate.

Johnson-Laird's suggestion that subjects leave the sentence represented in some uninterpreted form need not entail any addition to the range of representations available within the model. The blackboard (mental model) constitutes the sole medium of communication between the various modules that process linguistic input. As such it must already be representing the sentence under consideration, probably at a number of stages of abstraction between raw sense data and mental model. It would therefore seem natural to suggest that uninterpreted sentences would be held in one of those formats, particularly since sequences of meaningless words or syllables – such as telephone numbers and acquaintance's names – are regularly cached in memory. Moreover, since this process requires considerable effort, as shown by the difficulty of trying to do anything while keeping a telephone number in mind, the observed degradation in performance of the experimental task is to be expected. Curiously, however, Johnson-Laird does not pursue this path, but merely suggests that the uninterpreted representation of the premises employs an unspecified propositional format that is "close to their linguistic form". In the absence of any explicit motivation for postulating additional formatting, Occam's razor argues for specifying that the format used is one of those already employed for perception.

The essence of the language processing system Johnson-Laird proposes is the interpretation of a discourse by the construction of a single, unambiguous mental model. The experiments of (Ehrlich and Johnson-Laird, 1982) and (Mani and Johnson-Laird, 1982) explore how the system deals with a discourse that cannot be handled like this. They suggest the human language

[5] In both cases the references are to (Johnson-Laird, 1983) and (Mani and Johnson-Laird, 1982).

processing system resorts to attempting – not very effectively – to simply retain its actual linguistic input until the problem is resolved and it can be processed as normal. Moreover, the possibility of this kind of processing without interpretation allows the theory the room to begin to cope with a very fundamental feature of human language processing. As Fodor (1975, P159) points out, "Understanding is a graded notion, and it is possible to recover more or less of what a given utterance was intended to convey".

Johnson-Laird also points out that, not least because it provides only an incomplete description of a situation, language is necessarily ambiguous, which means that model building can never be done solely on the basis of the discourse. Instead, it must usually rely on making assumptions about unspecified features, and the experiments are only probing what happens when the relevant assumption generating procedure breaks down. Obviously, subsequent discourse may well contradict some of these assumptions, and Johnson-Laird proposes that this is normally dealt with by correcting the model being considered to make it compatible both with the new information and the original discourse. Of course, this can only be done if the relevant information is available – namely, the precise content of the original discourse, the retention of which Johnson-Laird is obliged to make part of the procedure for interpreting normal (incomplete) discourse.

However, this suggestion flies in the face of the evidence. Recall Fodor's throw-away demonstration during his characterisation of input modules, mentioned in section 2.1, of quite how quickly both syntactic details and the details of syntax are lost when text is read for comprehension. Moreover, if retaining uninterpreted information is a regular part of discourse comprehension, why did requiring the retention of only one proposition in Ehrlich and Johnson-Laird's experiment have such a detrimental effect on subjects ability to draw the situations being described?

The experimental evidence certainly supports Johnson-Laird's case for postulating the use of uninterpreted representations to deal with abnormal discourses, the case of normal discourse is much less secure. His argument is based on the notion of partially understood text, which he illustrates by offering (Johnson-Laird, 1983. P158-159) a thirty line extract from Conan Doyle's novel *Charles Augustus Milverton*. This describes how Holmes and Watson approach Milverton's house and break into it, and then details their progress as they make their way through it to his study. In the true fashion of a psychologist, he follows this text with an unexpected question concerning through which side of the building the pair forced their entry, a fact not explicitly mentioned in the text. He states, and experience bears him out, that few people can answer this question, although he claims that there is an answer implied by the information in the text. The reason, he suggests, is that the text is not fully interpreted into a model, which would have generated enough of the implications of the description to revealed the answer. Instead, it was only given partial interpretation, and the relevant detail was thus never computed.

However, this demonstration is most unsatisfactory. Johnson-Laird offers no explanation of **why** the text was given only a surface interpretation. There certainly is no reason to suggest that the text is not coherent or ought to be difficult to understand. Moreover, the account totally fails to fit with vivid experience of reading Conan Doyle's highly evocative text – evocative, indeed, of

powerful imagery, the subjective manifestation of the construction of a mental model! In other words, if reading this text does not evoke a mental model, no form of language will. Fortunately, there is an alternative explanation available, which should be familiar in the light of the tiger's stripes. The text presents the detectives' progress through a series of situations, each of which is interpreted by the construction of a detailed mental model of the situation described. Crucially, the answer to Johnson-Laird's question pertains to the inter-relation between them, and the limitation that only a single model can be entertained means that there is no capacity for capturing that relationship. Of course, when the text is re-read with the particular question in mind, this prevents "zooming" in on the details of each event, allowing a single global model to be created and the question answered.

On this approach, the surface form of normal discourse is indeed lost very quickly. Incompatibility between subsequent text and the model still causes modification of the model, but it must usually take place with only the benefit of whatever patchy record of the history of the discourse happens to be available in long term memory. Of course, this is far from saying that the modification will be made at random. On the contrary, it will be influenced by knowledge of the kind of situation being described It will be made on the basis of a lifetime of experience of repairing failed implicatures, and will take account of the fact that the problem with original model is (probably) the result of plausible (but erroneous) assumptions on the basis of the earlier discourse. Despite this, however, it is still perfectly possible that the modification made may be incorrect, and either contradict some earlier explicit statement, or inappropriately elevate some default assumption to the level of explicitly stated fact.

Finally, this latter possibility seems to be neatly illustrated by the answer Johnson-Laird gives to his question, which is not necessarily correct. It appears to depend on an assumption about the orientation of the corridor along which Holmes and Watson passed, or at least that there were no unmentioned branches from it. Without these assumptions, which Johnson-Laird seems to have made to facilitate the construction of his mental model of the discourse, it is possible to produce house layouts compatible with the break in having occurred at either side of the house. It would appear, then, that in this case at least, Johnson-Laird was unable to accurately check possible modifications of the model against the propositions specified in the text.

3.6. Models and Thought

So far, this chapter has attempted to show how the restrictions imposed on a cognitive architecture by phenomena related to mental imagery can be used to influence theorising in the domain of language processing. The rest of this thesis is directed towards illustrating its application to more general thought processes, as exemplified by syllogistic reasoning. However, this is a controversial task to undertake. There are a number of arguments that claim to show that, while they may have a very limited role in certain specialised kinds of mental activity, models can play no useful part in the processes of thought in general. Since the concept of a mental model is comparatively recent, they are mostly directed against mental imagery, which means they vary greatly in the difficulty of the challenge that they present to the richer information structure. Many of them merely demonstrate that a maximally impoverished kind of representation that is only

amenable to a very restricted range of transformations is insufficient to deal with some of the more subtle mental processes that are observed. As such, they have little bearing on the usefulness of a mental model. However, others reveal fundamental differences in the assumptions made about the inter-relationship of the processes supporting mental life.

One trivially countered argument, which it is hard to believe was put forward in earnest, is offered in (Fodor, 1975, P183). It is often suggested that it is crucial to the semantics of model based approaches that models (and images) resemble what they correspond to, or refer to, in the world. It is, of course, far from obvious what it is for an information structure in the brain, which is presumably embodied in some pattern of electrochemical activity, to resemble anything at all. However, since the perceptual systems are one of the main constructors of models, they define a mapping from situations in the world to models, and it is possible to suggest that "resembling" is intimately connected with the (or "a possible") inverse of it. In any event, Fodor tries to undermine the whole idea by asking his reader to study a drawing of a hexagon with a line linking each of its angles to the centre point. He then proclaims that doing so should have been a case of thinking of (or referring to) a cube, because his hexagonal figure "resembles a cube viewed from one of its corners".

However, in making this claim Fodor is (probably wilfully) overlooking the fact that interpreting a line drawing is a complex perceptual process. It takes place against a background of culturally established conventions, and its indeterminacy forms the basis of a number of famous illusions, such as the old-young lady, the duck-rabbit, and the goblet-faces. Indeed, it is possible to argue that in many respects the perception of a line drawing is more akin to language than vision. The existence of an alternative interpretation of a stimulus pattern is in no way reflected in the result of the perceptual process, which in this case is a full, multi-dimensional representation of the interpretation the perceptual mechanism has decided to adopt. The fact that a particular line on a piece of paper can be seen as either a duck or a rabbit merely illustrates, rather than undermines, the system's ability to represent these two objects differently. Moreover, their ability to **refer** is no more undermined by such an observation that the existence of the alternative parse as "ray gun" prevents a phoneme string naming the current American president.

Fodor also offers a more serious objection, which he traces back to Wittgenstein (1953). Lacking any clear characterisation of the properties of mental images (or models), he focuses on the fact that they must surely share many properties with other, more mundane, representations, such as photographs (or plastic toys). This seems reasonable, at least to the extent that the terms are well chosen, and leads to the suggestion that a mental image (or model) of a man walking up a hill should be thought of as much like a photograph (or diarama) of such an event. But he then asks what would distinguish this representation from that of a man walking backwards down a hill? Or hopping sideways round it? The "photograph" (or model) would be identical. He rightly points out that, if the system is to be seen as actually using the images, it could not be annotated in any way, for this leaves open the question of who reads the annotations; we are left with an undischarged homunculus, an obvious disaster for any theory of cognition. So how can such a highly ambiguous image be seen to represent anything?

Here, the solution lies in the fact that models are richer representations that photographs. While likening them rightly carries over the fundamental limitations that the two have in common, it has also carried over a merely coincidental technicality about photographs – that they record only static light intensities. It was argued in Section 1.2 that there is evidence to suggest that the representations involved are richer than this, and in fact the cine film and puppet theatre should be seen as providing better similes for mental images and models. Making this alteration – adding change with time – to our idea of a mental model immediately dissolves the problem. The man walking up the hill is clearly distinct from the walking backwards down it: He is going the other way! The problem is trivial once the representational resources available are extended beyond visual information (mental images) to include the offerings of the whole range of input and processing modules (i.e. mental models).

However, this ambiguity is only a small part of Fodor's attack, the remainder of which cannot be similarly deflected without embarking on a process of unprincipled extension of the representational power of the model. The main objective of Fodor's assault is to suggest the impossibility of using mental images (or models) to record a simple fact, like "John is fat". Once again, following from the likeness between a mental image and a photograph, Fodor suggests that a suitable representation would be like a photograph of John, showing that he was indeed fat. But, asks Fodor, why is this a representation of John being fat, and not, say, a representation of his having a beard, or equally not having a beard?

One possibility that the model theorist might be tempted to adopt the "trivial" solution of suggesting that the proposition be represented by modelling a piece of paper bearing the words "John is fat". Admittedly this would represent the proposition (and in a manner that will be useful later), but begs the question, which is vital for a theory of cognition, of who would read the modelled writing on the paper. In the absence of a suitable homunculus, this image/model is completely unable to capture the understanding of the proposition.

The alternative is to propose that models are made of parts – specific features are represented separately – and that it is possible to suggest some parts could be more prominent than others. This would make it possible to have a model in which certain properties could be made particularly salient, by in some sense "highlighting" the primitives that depict them. Thus the proposition that "John is green" would be represented by a model which, if described in the formalism of frames (Minsky, 1975) might look like

```
(schema   GOO34
          (is-a person)
          (name John)
          (weight 12-stone)
   >>>    (COLOUR GREEN) <<<
          (age 25))
```

It is certainly reasonable to suggest that the information structure that constitutes a mental model is an aggregation of a number of separate sub-structures which are to some extent independent. Indeed, there is ample evidence that certain aspects of a model can readily be omitted altogether – that is, people can undeniably entertain partial representations of situations.

This is clearly illustrated by everyday activities like identifying a present through its wrapping or simply seeing something on a monochrome television.[6] These actions clearly involve the formation of a multi-modal representation of an object or situation that can certainly interact with stored knowledge on a variety of topics – they are certainly mental models. Moreover, they can obviously proceed despite the complete absence of certain kinds of information – such as the colour of the object involved. Thus it seems reasonable to suggest that, if colour can be omitted, so too can it be emphasised, and by such emphasis can "John is green" be represented.

There are, however, limitations upon the generality of this mechanism. This structure within the information – the inter-relations between things that can be omitted – is part of the characterisation of the code in which the model is represented. As a result, there is necessarily a finite range of degrees of freedom – "primitive" dimensions on which the situation being modelled is evaluated which, if the contents of the model is to be manipulated sensibly by processing modules, must change only slowly, if at all. There is evidence to suggest that the colour of an object constitutes such an independent feature, which is why the example of "John is green" seems so reasonable. However, it is not possible to use "highlighting" to represent an arbitrary proposition, because in general there will not be a suitable "slot" to emphasise. The schema for John, as derived from that of a typical person, simply will not have anything that can be highlighted to capture "John is heavier than a hamster". Since holding a belief is a matter of standing in a computational relationship to a representation, and mental models cannot be used to represent even the simplest of beliefs, it appears that they are therefore quite unable to account for cognition.

There are two fundamentally different ways in which a theorist who supports the idea of mental models can react to this. The first, and probably the most obvious thing to try, would be to simply answer Fodor's claim by showing how a single model could indeed represent a proposition. Such an approach would begin by pointing out that a model is only a model in view of the fact that there exists an associated set of procedures for interpreting and manipulating it. The theorist could then start to stipulate conditions on the model-interpreting routines that would restrict the range of things that the model could be seen as representing.[7] Thus, by restricting the things that the procedures could "read off" the model, the theorist could doubtless manage to produce a mechanism whereby a suitably chosen "photograph" of John is seen to represent the fact that John is fat.

However, it is far from obvious that this is a good move for a theorist trying to establish the usefulness of mental models. The fact that logicians' models are characterised by nothing more than the set of propositions that are true in them suggests an obvious sense in which any other model can be similarly regarded as a collection of propositions. Indeed, images (and therefore presumably models too) are readily represented in computers, in essentially propositional form.

[6] It is also central to any attempt to involve models in the representation of the intentions or unique knowledge of others – See (Reichgelt, 1986) and (Johnson-Laird, 1983. P430 ff).

[7] This is effectively the step taken by Johnson-Laird in order to allow conceptual mental models to support specific conclusions to syllogisms – see chapter 6.

Thus the onus is on the mental model theorist to show how introducing the concept has any advantage over simply talking in terms of inferences and propositions. It is not obvious that this end is best served by suggesting how a model can be seen as equivalent to a proposition! When Fodor points out that a model cannot represent a proposition, the mental model theorist should not oppose him, for all he is doing is highlighting the fact that the distinction between propositional and model based systems of representation is not vacuous. If, by suitable ingenuity, a theorist managed to produce some means by which an image or model could be seen as representing a proposition, then he would have gone a long way towards undermining the justification for postulating the image in the first place.

Where does this leave mental models? Fodor has produced a challenge to theorists who would try to use the concept of a mental model to enrich their beliefs about cognition. He believes it to show that mental models are fundamentally incapable of explaining even the simplest of facts about mental life – that people can be said to have beliefs – and are thus totally useless as the foundation of a paradigm within which to approach the study of cognition. Yet the mental modeller is obliged not to even try the most obvious response by attempting to show how Fodor's objection can be met. If a theorist is to support the idea of mental models as a useful theoretical concept, some way must be found to dispel the power of his arguments. What is the alternative for the mental modeller?

The solution lies in realising that Fodor's deduction rests on a false assumption. It does not follow from the fact that holding a belief is the state of standing in a computational relation to a representation that the representation involved must be of that belief. This means his problem need not be met head on and answered, but can be merely side-stepped and dismissed as irrelevant. The mental modeller can cheerfully admit that a mental model cannot represent a proposition, but insist that this is not important because people do not represent propositions either. For, although there may be overwhelmingly evidence for the presence of a belief, in the form of behaviour that can be described as motivated by it, none of it is relevant to the existence of a particular representation of it.

To see this, consider the kinds of evidence that would normally support an ascription of a belief. One would obviously be the subject's direct expression of, or assent to, the belief. For instance, one might be inclined to say that someone believes that John is fat it they were to say "I think John is fat", or offer an affirmative answer to the question "do you think John is fat?". But people only ever explicitly express a very small proportion of the things that we would freely describe them as believing, so explicit expression can certainly not be seen as a necessary condition for belief ascription. A more common type of evidence is the behaviour of the subject. A belief that John is fat could be deduced from a variety of observed, possibly verbal, behaviour. This could be something very direct, such as referring to John as "the fat guy with the beard" or making jokes about his weight or size. Or it could be more indirect, such as suggesting that John should be told about a new outsize clothes shop that has opened, or suggesting that the seating for a dinner party be so arranged that John is not given a delicate antique chair. Finally it is common to ascribe beliefs to others on the basis of one's own opinions of the world. One can assume that

Fred believes that John is fat is Fred knows John, and John is pretty undeniably fat.

It should now be clear that the mental model approach can ignore Fodor's arguments if it can offer the hope of a system that can produce the ascription-motivating behaviour without needing an explicit representation of a belief. This "system" can be distributed throughout the postulated cognitive system, with all the different sorts of behaviour that would support the belief potentially coming from separate sources. In the absence of a single representation of the proposition, the grounds for supposing that it is believed arise from the correlations which appear between these independent behaviour mechanisms. However, it is no mystery that the mechanisms correlate in just such a way that we can capture it by means of an ascribed belief. The human conceptual and linguistic systems have evolved in the presence of just such reliable and predictive correlations, and they are built for the job.

It might help to consider how this applies to the now-familiar example of John's alleged obesity. Verbal behaviour related to this belief, such as asserting or assenting to the proposition, would occur because the internal "mental model" of John (i.e. the template used to generate representations of him in mental models of situations involving him) fitted the criteria that allow the system to use the word fat – a property of the system's linguistic abilities. On the other hand, the desire to tell John about the outsize clothes shop would be seen as arising because the (mental model of the) customers that the shop caters for has similar physical dimensions to the (mental models of) John. This would be completely unrelated to the language abilities, and would constitute an entirely different behaviour producing mechanism, quite independent of the word (or explicit concept) fat. The only connection between the two sources of behaviour is that both use the same representation of John (although the non-verbal processes need not make any reference to him by name, or even be capable of conceiving of his having a name).

Moreover, it is not only possible to avoid assuming that ascribing a particular belief involves a commitment to its explicit representation, it is necessary. People are obviously capable of holding an infinite number of beliefs, and indeed all do, under the label of "common sense". For instance, suitable empirical investigation (e.g asking him) might give rise to evidence that justifies the hypothesis that Frodo believes that a particular leaf weighs less that a hamster. Yet on the basis of this one unproblematical belief it is possible to derive infinitely more, each every bit as plausible and innocuous (not to mention peculiar) simply by increasing the number of hamsters in the comparison weight. Obviously this infinite collection of beliefs cannot be represented in Frodo's necessarily finite brain, which means that it is a clear case where beliefs should be ascribed even though definitely not represented. Of course, it is possible to criticise this example as a "trick" that relied on the use of a formalism (arithmetic) to generate an infinite number of "apparent" beliefs from one "real" one. In other words, the beliefs were not really independently held at all, but were somehow worked out using the logical properties of arithmetic and a single explicit representation of the "strongest" belief. Clearly, something like this must be true, and the explicit propositionalist has only been forced to give a little ground in the area of mathematical consequences.

However, there is considerably more mileage in this example, since it is possible to generate an (effectively) infinite number of (again innocuous, plausible and peculiar) beliefs just by taking Frodo to the nearest forest in autumn. Now those who would suggest that all beliefs are represented are forced to make a more significant retreat, in that they can no longer rely on the explicit and in some sense necessary rules of arithmetic to get them off the hook. There is no necessary relationship between the objects blowing about on the forest floor. There is no straightforward way in which they can be seen to be related to each other – they will differ in size, shape, colour, age, weight, place of origin, composition, and indeed every physical parameter that can be detected by human (and probably any other) senses. Thus there is no obvious "base" belief, from which the others can be said to follow. The relevant belief ascriptions can only be justified on the basis that the evidence available indicates that should he ever be in a situation in which the relative weights of any of the leaves and a number of hamsters was relevant, Frodo would behave in a manner consistent with the belief being ascribed.

Dennett (1978) highlights this same weakness in Fodor's theory.

> Perhaps we "entertain" propositional attitudes either seriatim or at least in manageably small numbers at any one time, but the propositional attitudes we **have** far outstrip those we (in some sense) actively entertain. For instance, it should come as no news to any of you that zebras in the wild do not wear overcoats, but I hazard the guess that it **hadn't occurred** to any of you before just now. We all have believed it for some time but were not born believing it, so we must have come to believe it between birth and, say, age fifteen, but it is not at all plausible that this is a hypothesis any of us has explicitly formed and confirmed in our childhood, even unconsciously. It is not even plausible that having formed and confirmed other hypotheses entailing this fact about zebras, we (in our spare time?) explicitly **computed** this implication.
>
> (Dennett, 1978. P104)

He goes on to suggest that if believing a proposition must involve being in a relationship to an explicit representation of it, "being disposed to represent" must count as such a relationship.

It is important to notice that Fodor's commitment to the explicit representation of beliefs is not an isolated feature of his theorising, and Fodor (1975) shows that it is intimately connected with his overall view of mental, and in particular language, processing. He argues that cognition is supported by a processing system that operates by the making of inferences based on propositions expressed as sentences in a "Language of Thought". Such sentences present within the brain constitute the beliefs of the system, and they are processed in a way that honours their semantic import. Moreover, it is the form in which the sensory systems – the input modules – offer their views of the environment, and in which information is amenable to combination with the system's knowledge of the world. In this it is clearly playing a very similar role to that of the mental model, and indeed it is tempting to suggest that a model is simply a collection of sentences and that processing modules are simply additional inference engines with access to specialised inference rules. However, there are certain features of the Fodor's system, particularly apparent in the way it handles language, that violate many of the essential features of the model approach.

Fodor (1975, P103-110) characterises language as a number of shared conventions for facilitating the exchange of *messages* by providing mappings between them and acoustic wave forms. These messages take the form of a sentence within the language of thought which contains the information communicated. Thus the successful use of (for instance English) involves two

stages. The speaker must formulate a message in the language of thought and attempt to produce a wave form that satisfies the criteria of one of the specifications that the conventions of English indicate should be used to communicate such messages. The hearer must perceive that wave form as satisfying those criteria, and thus by use of those same conventions gain access to a language of thought representation of what the speaker wanted to say. Crucially, to do so it relies on the speaker being able to formulate an explicit expression of the precise message to be transmitted. Similarly, the hearer must be equally able to extract that message from the waveform and couch it in a form that enables its entailments to be subsequently derived using knowledge expressed as meaning postulates. In other words, Fodor assumes the psychological autonomy of intentions thesis that Johnson-Laird rejects.

This direct translation is no isolated quirk of Fodor's theory, but is central to his views of the process of understanding a sentence and leads him to explicitly reject one of the central tenets of the model-based account.

> It should be borne in mind that understanding a sentence involves computing a representation of the sentence that **determines** its entailments; it doesn't involve computing the entailments.
>
> (Fodor, 1975. P150)

Thus Fodor explicitly separates the generating of the Language of Thought translation of the message of a sentence – which he identifies with understanding it – from exploring its interaction with previous knowledge of both the discourse and the world. Given his linguistic view of the format of inner representations, he characterises this exploration as at least involving the replacement of complex terms by their definitions, and justifies his decision by suggesting:

> In short, we have two broad theoretical options: We can acknowledge definitions instead of meaning postulates and thereby simplify the logic at the cost of complicating the sentence understander, or we can acknowledge meaning postulates instead of definitions and thus simplify the sentence understander at the cost of complicating the logic. The present point is that, ceteris paribus, we would be well advised to go the second route. For the important thing about sentence understanding is that it is **fast**; too fast, in fact, for any psycholinguistic theory that is currently available to explain. We make this mystery worse in proportion as we make the relation between wave forms and messages abstract, since it is this relation that the sentence understander is required to compute.
>
> (Fodor, 1975. P151)

Moreover, having established this separation, he uses it to motivate an explanation for the familiar observation that "understanding is a graded notion" (ibid, P159). He suggests that this arises from the fact that the inferences needed to determine the consequences of a sentence require computational effort. Since any realisable system will have finite limits on available computational power, there must come a point where the search for further consequences must be discontinued, even though there may be others that would be accessible to further processing.

This is, of course, quite unlike the characterisation of the process of sentence comprehension assumed by the mental model approach, and the facts of language processing are hardly supportive of Fodor's position. Consider the following opening to a telephone conversation:

> Hi Jerry, its Phil!
> The local car thieves visited College Street last night, so I can't give you a lift in this morning. My car has got no wheels!

For Fodor, understanding this discourse requires a representational system that is able to explicitly

represent propositions in order to capture the message of each sentence. In particular, it must be able to suitably manipulate expression explicitly representing the negation in the last sentence.

Within an account based on mental models, a sentence is not seen as just an encoded statement of a fact, to be represented for future contemplation. Instead, it is more like a recipe for appropriately updating the intended hearer's ideas about the situation under discussion, an idea that accords well with the notions of Wason (1965, 1972). Thus the particular case of the last sentence in the above discourse is better interpreted as an instruction to Jerry to modify his ideas about Phil's car. Since he was apparently expecting to be given a ride in it, his current representation of it probably describes it as having wheels, and his understanding the sentence amounts to making such alterations as will change this. The result of doing this will contain no negative information, which is just as well, since mental models are characterised as being unable to handle negation. Instead, it will be a description of an object that is like a car in every respect, except that where a car has wheels, it has piles of bricks, or simply hubs resting on the ground in empty wheel arches. Of course, the discourse does not specify which of these is appropriate, but merely invites the hearer to make assumptions on the basis of his knowledge of (or guesses about) the way thieves leave cars when they steal the wheels. Facilitating such supporting inferences on the basis of (necessarily) incomplete discourse is the central contribution of a mental model to language processing. If the resolution of the ambiguity is considered sufficiently important, the model may serve only to highlight it and thus trigger a question to clarify the matter.

Similar arguments can be advanced to show that a theory based on mental models can also account for the production of negated sentences from internal representations that contain no negation. Thus if Jerry chose to describe Phil's car in the condition in which he now conceives it to be, he would discover as a result of examining the range of noun phrases open to him that the noun that came closest was "car". However, the thing to be described is significantly different from the prototypical car, in that it is entirely deficient in terms of wheels. Thus he would describe it as "a car with no wheels", or "a car with its wheels missing". Once again, however, there would be no negation represented in the information structures supporting this use of language, which simply record the way entities in the world are, not what they aren't.

The desire to reduce the power of the underlying representational system is not the only reason to disagree with Fodor's position and accept that "discovering the entailments" of a sentence is part of the process of understanding it. The information most people immediately extract from the last sentence in the suggested telephone conversation is that the wheels of Phil's car were stolen last night. Despite any logicians' tendency to argue that its truth conditions are compatible with Phil having sold them to the car thieves, there is a very real sense in which that sentence is not only a description of a state, but also of a theft. Of course, Johnson-Laird's whole point is that recognising this necessitates taking account of the semantics of the words in the sentence, and doing that is nothing less that "discovering the entailments" of the sentence as part of understanding it.

This example far from exhausts the extent to which world knowledge penetrates the interpretation of this discourse, as is shown by considering Phil's other use of negation in his

telephone conversation. In the third sentence he states that he is not able to give Jerry a lift. It should be obvious, in the light of the above discussion of wheelless cars, that this should be interpreted as a command to Jerry to remove his expectation that he was going to be taken somewhere in Phil's car. This is fine, provided that we make the assumption that Jerry was expecting such an event: The more puzzling aspect is how the discourse can be understood by anyone else. It is hardly obvious how the denial can be interpreted as a removal of an expectation when most people encountering the discourse had no preconceptions about Jerry's intended travelling arrangements. Nevertheless, the discourse makes perfect sense to anybody.

The problem arises from the fact that the negation refers to – i.e. calls for the cancelling of – an intention of which the hearer has no knowledge. However, the use of bridging inferences to introduce a new feature into a model in order to interpret a sentence that mentions it has already been identified as central to discourse comprehension. In this case, the new feature is not a physical object, but some kind of expectation, on the part of Jerry, that Phil was going to give him a lift somewhere. The most likely way that this could have come about would have been if they had arranged the lift, so it is reasonable to assume that this has indeed happened some time in the past. Of course, since people rarely arrange lifts for the fun of it, this in turn implies that Jerry is wanting to be taken somewhere. Hence when someone is confronted with this discourse, they construct a mental model of not only a telephone conversation and a vandalised car, but also of a desire on the part of Jerry to go somewhere.

Finally, it should be emphasised that Fodor's argument that minimising the requirements for enormous processing speeds by deferring the generation of the implications of a sentence is of very limited use. It is an empirical fact that many of the implicit inferences underlying a complex sentence are available more or less immediately – the role of Phil's last sentence is really only to specify precisely what the vandals have done to his car. Thus Fodor's ploy of moving the generation of these consequences from language understanding to general reasoning doesn't reduce the amount of work that must be done extremely quickly. It merely transfers the responsibility for doing it from the language processing input modules to the engine responsible for general thought. Unfortunately, while the former is already known to operate at awesome speed, the latter, as Fodor (1983, P63) himself admits, is very much slower. As a result, it is hard to see the merit in the suggestion.

Moreover, to retain abstract, uninterpreted terms in the manner that Fodor suggests would be to rule out one of the most exciting processing architectures yet discovered – namely, one involving multiple, specialised, processing agents. Such agents could only cooperate effectively by using a common code to communicate, and only by expanding complex terms into primitive components in that code could their contribution to the situation be made available to them all. Within such a framework, partial understanding is given a very different characterisation, which will be mentioned again in Chapter 8.

3.7. Abstract Thinking

The discussion above has sought to show that the mental model can indeed be seen as a useful notion to employ when considering the mental representation of certain kinds of knowledge about

the world. However, it is possible to counter that it only demonstrates an ability to handle very limited, simple types of information, such as physical situations and scenes. With the exception of the brief discussion of the Symbolic Distance Effect in section 3.2, no suggestion has been made of their relevance to any more abstract situations. Indeed, no account has even been given for the representation, within the proposed restrictions on specificity and negation, of abstract knowledge such as "Debra likes strawberries and cream" or "Phil has insured his car with Legal and General". These things are very easy to represent in a propositional system, and yet there is no obvious extension of any approach that alludes to Airfix kits and photographs to encompass such matters. Thus, there is still work to be done before mental models can be seen as a viable tool for the description and explanation of cognitive activities in general.

One of the most fundamental problems concerns the range of features that a model can represent. As was pointed out in Chapter 1, it is impossible to consider the contents of an information structure in isolation from the processes that manipulate it. Nonetheless, it is possible to seek to characterise it by analysing the performance of the combined system in terms of its relative difficulty in dealing with (or omitting) features of a situation, the relative accessibility of information and the kinds of deterioration patterns it exhibits. On the basis of such evidence it is possible to infer something of the contents of the information structure, and given the central role being assigned to mental models, the characterisation of their representational power is obviously of some concern.

Johnson-Laird suggests three conditions on the range of things that can be represented within a mental model, which vary in both their importance and their plausibility. He couches these in terms of the representation of "concepts", but he does so without clarifying what he means by this (in which, of course, he is far from alone). Certain aspects of his discussion seem to be alluding to the dimensions of analysis along which the model characterises its contents – the "slots" in the schema – while others seem to be pitched at quite a different level.

His first condition is

> The predicability principle: one predicate can apply to all the terms to which another applies, but they cannot have intersecting ranges of application.
>
> Thus, for example, "animate" and "human" apply to certain things in common, "animate" applies to some things to which "human" does not apply, but there is nothing to which "human" applies and "animate" does not.
>
> (Johnson-Laird, 1983. P411)

This is a curious suggestion, flying as it does so squarely in the face of so many obvious apparent counter-examples. Johnson-Laird admits their existence, and tries to suggest that they often only appear as such because of the lack of a "uniform compositional principle for the interpretation of adjective+noun constructions" (ibid) which allows the situation to be complicated by "derivative" uses of adjectives. Unfortunately, many of them, such as "edible fungus", "dangerous animal" and "monosyllabic noun", resist such attempts.[8]

[8] It is possible that Johnson-Laird is actually putting forward an argument that would hold given a careful distinction between a predicate that *holds*, in the sense of "is true of", and one that *applies*, in the sense of "it is sensible to consider whether...". If this is the case, Johnson-Laird's illustration using "human" and "animate" could have been better chosen. Moreover, there is considerable work to be done to explicate the direction of which Johnson-Laird does not even hint at.

Johnson-Laird's second condition – his *innateness principle* – is that "all conceptual primitives are innate", though unfortunately it is not clear what this is intended to convey. Any system without innate primitives would have to use only acquired concepts, and as Fodor (1975) argues, it is far from obvious how concepts can be acquired by a system that has no representational apparatus. Therefore, every system must have **some** concepts which it is innately able to represent. Equally, however, adult human beings clearly have certain concepts which are entirely cultural in origin – Johnson-Laird himself suggests Mrs Thatcher and the H-bomb – and thus by no stretch of the imagine innate. This means that his use of **all** must reflect the fact that Johnson-Laird is dealing with the origins of only the system's conceptual **primitives**. The innateness principle, therefore, seems to amount to a rather obscurely phrased definition of a conceptual primitive – a suggestion that all the "slots" in the schemata are innate.

Finally, Johnson-Laird's third restriction is that there are a finite set of primitives, and a finite number of ways of combining them. For this restriction to carry any weight – there are, after all, only a finite number of atoms in the earth – "finite" must be interpreted as "small enough to hope to discover, or at least be influenced by the presence of, the upper bound". He then attempts to characterise both the range of the innate primitives and the ways of combining them by illustrating the extent to which English words can be defined in terms of a small subset of concepts. However, the importance of this demonstration to the proposal concerning the representational resources available within a mental model depends on the closeness of the mapping between linguistic forms and internal representation. Thus he is tacitly assuming that (certain) natural language words have simple definitions in terms of mental models, an assumption that requires considerably more justification than it receives.

Despite these problems and uncertainties with Johnson-Laird's attempts to do so, there is no doubt of the importance of characterising the possible content of a mental model, and, in particular, the extent to which it can be enhanced by experience. Any particular concept is a way of grouping together a number entities or situations that can in some way be beneficially treated as alike. Within the proposed cognitive architecture, having a concept can been characterised as involving possessing processing modules both to recognise and to behave appropriately towards instances of it. This means that learning a new concept is closely related to developing new processing modules. The evidence presented in Chapter 2, from Stamm's electrical disruption of the object concept, suggests that the mental model (blackboard) carries information from such acquired processing modules. However, it is worth considering whether this might be by extending the code of the model – adding a new type of slot to the schemata that characterise it – or by using new (or otherwise inappropriate) patterns of existing features?

Considering this question after an encounter with a life insurance salesman, an art critic or an economist would probably lead one to suggest that people can certainly enrich their representation of the world with fundamentally new concepts – that is, it is possible to learn new dimensions along which to evaluate and think about situations. However, the restriction on the communication between modules is one of the strongest constraints on the power of a blackboard architecture, and the ability to relax it by adding new representational primitives would constitute an enormous

increase in the power of the system. This means that the psychologist should strive to avoid postulating that concept acquisition requires this – which is in line with the suggested interpretation of Johnson-Laird's innateness principle. This is clearly straightforward in the case of processing modules that are closely related to physical states of the world, such as the object concept.[9] In other cases, however, it is much less clear how this requirement should be met, perhaps to the extend that it is even tempting to take the proposal as something of a straw man. However, this would be a mistake, since if an architecture that cannot extend its representational resources is able to cover the data, postulating it constitutes the best possible theory.[10]

It is important to recognise what is likely to be a necessary tradeoff between innate representational primitives and the ability to create new ones. To avoid needing to enrich the representational vocabulary, it is necessary to ensure that it is (innately) rich enough already. This can be illustrated by considering the results of DeGroot (1965, 1966). He describes experiments exploring subjects' ability to recall an arrangement of pieces on a chess board after only brief (5 second) exposure. He reports that while chess playing ability is of no help in the case of random arrangements of pieces, good players are very much better at reconstructing positions taken from games – i.e. where they form the kinds of patterns that are important in chess. Clearly, recall is being affected by the ability to detect, within 5 seconds, these significant relations between pieces.

The most obvious way of interpreting these data is to suggest that the relevant chess-related relations, such as "can capture", are being added to the representation of the situation that is being committed to memory. This implies that acquiring mastery of chess involves (among other things) the development of a richer representation of (chess-related) situations. More generally, it implies a cognitive architecture that can extend the representational resources available for communication between modules when required.

There are, however, alternatives that do not require this flexibility. One is to deny that the chess-specific relations are communicated at all, and to take advantage of the dynamic nature of the model and processing modules to suggest that they remain local to the relevant processing module. Instead of signalling its recognition of a particular interaction by enriching a static representation, the relevant module could reveal the results of its analysis of the situation by attempting to indicate its likely development. In this case, chess mastery would improve performance because the specialised processing modules would react differently to incorrectly remembered situations. Of course, this requires that a task-specific processing module be able to execute (and signal the result of) a comparison between its evaluation of two different situations – that it must, at the very least, be able to deploy some kind of memory. This is, of course, an extension of the cognitive architecture. However, it is a less powerful one than adding the ability to add relations when

[9] Though of course this is only the case because the two-dimensional arrangement of surface information proposed to account for mental imagery has already been replaced by a richer, 3D representation of a state of the world.

[10] It is tempting to draw an analogy with Newton's first law of motion, which states that unless disturbed, physical objects will move at constant speed in a straight line. Such a suggestion is immediately confronted by an enormous array of apparent counter examples: something released in mid-air falls, sliding blocks stop, leaves swirl and pendula swing. Yet despite the fact that this ideal behaviour can at best be observed only momentarily, the history of physics shows that it provides a firm foundation for understanding mechanics.

required, and, moreover, constitutes a highly plausible feature of a self-contained (modular) processing system.

An alternative suggestion is that the relevant processing modules can record and communicate their chess-related information by the use of some general, "attack", "threaten" or "protect" relation that is already present for some other purpose. In doing so, they would be using existing representational primitives in a novel (because otherwise inappropriate) situation – i.e. between pieces of wood. Of course, this presupposes that the ability to represent these abstract relations within models is available, which is, of course an extension of the architecture supporting them. However, once again there is a modicum of evidence on which to base a case for the innateness of the concepts involved. In a world where predators exist, an awareness of danger, together with the threat of possible danger, must be used to shape every aspect of daily life. Indeed, the ability to recognise (some forms of) danger and threats to self is certainly innate, and animals appear well able to extend this both to their young and to other members of their social group. Moreover, the suggestion that acquired concepts utilise existing representational resources offers a natural explanation for the existence and importance of metaphor. If the representations of newly acquired concepts constitute elaborations of those already employed for other purposes, they will resemble them. That is, they will (partially) match the patterns used to direct the manipulation of the existing concept and, in particular, its linguistic expression.

In the light of this discussion of the desirability of not enriching the range of features that the model is postulated to represent, it is possible to return to the matter of how a model might represent such abstract propositions as "Debra likes strawberries and cream". Before doing so, however, it is worth at mentioning how a "propositional" system tackles such matters. For instance, Fodor might suggest that Debra's desire might be captured in the Language of Thought by a sentence such as

likes(Debra, strawberries-and-cream).

However, this is far from ideal, not least because the representation does not capture the fact that the object of the desire is made up of components, and shares properties with them. This could readily be changed, to something along the lines of

likes(Debra, Combination34),
involves(Combination34,[strawberries, cream]).

Even if this is done, however, the real kernel of the problem remains. The meaning of the proposition is represented only to the extent that the meanings of the predicates "likes" and "involves" are known. The propositionalist is forced to rely on a large range of predicates – virtually as large, as Fodor has argued, as the vocabulary of English. Thus although he can cheerfully represent all kinds of knowledge, he can do so only at the expense of relegating all interest in the understanding of the world to the mechanism by which these propositions are manipulated. The baby has been thrown out with the bath-water, and the problem of how we understand language and the world about us has been answered by saying that we are innately able to do it. The fact that the mechanism is innate makes no difference to the importance of the question of how it works, and that is precisely the one that Fodor is begging.

In considering how a model based system can deal with abstract beliefs or knowledge, it is important to recognise and reject an assumption that has crept into most psychological theorising. It is generally accepted that most human knowledge resides in a "long term" memory, of effectively limitless capacity but sufficiently difficult to manipulate that thought processes are restricted to a much smaller "working memory". It is usually assumed that this working memory is a holding area for a passive representation of a situation, and that thinking consists of manipulating its contents in line with the knowledge of the world stored permanently in long-term memory. However, the blackboard is not a static snapshot of a scene, but a rich information structure that is being continually monitored and updated by a battery of processing modules. As such, it can hold information not just in its state, but in the ways it will change state under the influence of processing modules. This adds literally another dimension to the representation of abstract concepts, in that such knowledge can be stored within processing modules in the form of rules for transforming models.

To give a concrete example of this, knowing that Debra likes strawberries would be a matter of having a processing module that would predict that in any situation in which Debra has the opportunity to get some strawberries, she would attempt to obtain and eat them. In order to extract this knowledge from its inaccessible resting place, the cognitive system would have to access it indirectly, via the blackboard. Answering a question about whether Debra liked strawberries would involve depositing upon it a description of a situation involving Debra and strawberries and observing what happened. Such a model might serve to key the recall of a specific event in which Debra was confronted with some strawberries, in which case the answer could be given on the basis of how she reacted. But equally it might simply result in Debra being represented as wanting to get at the strawberries (recall that at least limited intentions can be represented directly).

This is, of course, only a single instance of the indirect representation of an abstract concept. As such it serves only as a pointer to the **kind** of mechanism that may allow such notions to be dealt with within an architecture with the kind of restrictions that have been argued for thus far. The next four chapters are concerned with the mechanisms by which subjects with no training in logic tackle a specific reasoning task – solving categorial syllogisms – which involves handling the abstraction of quantification. In particular, Chapter 4 describes the task and discusses the meta-theoretical constraints upon a suitable theory, while Chapter 5 outlines the traditional theories offered to explain subjects' performance, together with the experimental results that motivated them. Chapter 6 both describes Johnson-Laird's experiments, and presents and criticises the account of syllogistic reasoning using an elaboration of mental models which they led him to formulate. Finally, Chapter 7 proposes a new account of the processes involved in syllogistic reasoning which uses mental models that satisfy the proposed constraints.

CHAPTER 4

The Syllogistic Reasoning Task

4.1. Introduction

The study of the syllogism began with Aristotle, who thought that all deductive reasoning was syllogistic, and that if only all the valid syllogisms could be stated, fallacies could be avoided. Despite the shortcomings of Aristotle's system (Russell, 1946. Chapter XXII), it became the backbone of logic from his time right up until the end of the nineteenth century, and was still an important part of the logic syllabus when Wilkins began to study the reasons for logic students' poor performance in 1928. Other psychologists also adopted the syllogism as a medium for studying reasoning, and Johnson-Laird highlights its merits, pointing out that the same simply stated problem format encompasses problems that range from trivially easy for almost everyone to almost impossible for almost anyone. As a result, there is a substantial body of literature on syllogistic reasoning.

This first section of this chapter describes the categorial syllogism and presents the issues underlying logically acceptable performance. The second section discusses why it should be a focus for psychological investigation and the third section suggests what issues are relevant to judging the worth of a psychological theory in the area. Since subjects are poor at solving syllogisms, dealing with their reasoning errors is a central feature of any theory, and the fourth section attempts to distinguish the different kinds of ways that these can be explained.

4.2. The Logic of Syllogisms

4.2.1. What are Syllogisms?

Consider a statement of the form "some whatsits are doobries". One of the ways that it can be interpreted is as expressing a relation between sets of individual entities: In this case, that within a certain domain, the intersection of the set of whatsits with the set of doobries is not empty. Similar readings follow naturally for related sentences obtained by replacing "some" by "all" or "none", or "are" by "are not". This gives four possible sentence "shapes", or *moods*:

"All of the whatsits are doobries"	(known as **A**, from Affirmo)
"Some of the whatsits are doobries"	(known as **I**, from affIrmo)
"None of the whatsits are doobries"	(known as **E**, from nEgo)
"Some of the whatsits are not doobries"	(known as **O**, from negO)

Henceforth, sentences in these forms will be described by quoting their mood and the predicates involved: e.g. the first sentence above will be described as **A(whatsits, doobries)**.

If two such statements are considered then, provided they have a noun phrase, or qualifier, in common (known as the *middle term*), it may be possible to combine them to deduce a third sentence of similar form. Thus, given the premises "All of the whatsits are doobries" and "All of the doobries are thingies", one can deduce that "All of the whatsits are thingies", and one can do this without any idea of what whatsits, thingies of doobries might be.[1] The combination of such premises is the basis of the categorial syllogism. With two premises, each in one of the four forms, there are 16 possible combinations.

In addition to the forms of the individual premises, there is another degree of freedom within the syllogism, namely the order of the predicates within the premises, and in particular the position occupied by the middle term within each premise. Thus for the case of what will henceforth be taken as the canonical syllogism, relating As and Cs with Bs as the middle term, there are four possible arrangements of the premises:

1.	2.	3.	4.
A - B	B - A	A - B	B - A
B - C	C - B	C - B	B - C

Following Johnson-Laird, this will be referred to as the *figure* of the syllogism, with the numbering shown above serving to name the figures.

This numbering of figures must be approached with caution, since although it closely resembles the traditional use of the term, there are important differences. Traditionally, the figure of a syllogism includes not only the premises but also the conclusion, which always draws its grammatical subject from the second of the premises. Thus in the example above, the conclusion could only quantify how many Cs were As. If the order of the terms within the conclusion were reversed – a process known as *conversion*, and yielding the converse of a proposition (e.g. "Some Cs are As" is the converse of "Some As are Cs"[2]) – the resulting syllogism could only be accommodated into one of the traditional figures by reversing the order of the premises. While clearly this has no effect on the logical properties of the situation, which explains why the traditional figure system is employed for most work with syllogisms, its (potential) effect on the **psycho**logical properties led Johnson-Laird to adopt his alternative system.

Taking both the factors of figure and premise form into account, the fact that there are 64 possible syllogistic problems, 16 in each of four figures, should be reasonably obvious. Provided it is clear that Johnson-Laird's numbering is being employed, combining the number of the figure with the letters denoting the mood of each premise allows syllogistic problems to be identified. Thus **2EI** refers to the syllogistic problem in the second figure, with the first premise in **E** form and the second in **I**. Thus its premises can be written as **E(b, a) I(c, b)**, and an actual deduction of this form would be:

2EI None of the beekeepers are artists.
Some of the chemists are beekeepers.

---

[1] I.e. **A(whatsits, thingies)** follows from **A(whatsits, doobries)** and **A(doobries, thingies)**.

[2] I.e. **I(c, a)** is the converse of **I(a, c)**.

4.2.2. The Right Answers

Having considered the nature of the problems, it seems natural to consider the solutions. A valid conclusion is one that must be true in any situation in which the premises hold, and Table 4.1 shows the possible premise pairs together with their valid conclusions. Neglecting the possibility of already knowing all valid conclusions, the solving of a syllogistic problem thus involves the task of discovering whether a particular conclusion is true throughout the range of possible situations permitted by the premises (although there is no a priori necessity for the task to be viewed as such). This range arises from the combination of two factors.

A(a, b)	**I(a, b)**	**E(a, b)**	**O(a, b)**	**PREMISE FORM**	**A(b, a)**	**I(b, a)**	**E(b, a)**	**O(b, a)**
A(a, c) I	I	O(c, a)		**A(b, c)**	I	I	O(c, a)	O(c, a)
		O(c, a)		**I(b, c)**	I		O(c, a)	
E O	O(a, c)			**E(b, c)**	O(a, c)	O(a, c)		
				O(b, c)	O(a, c)			
A(a, b)	**I(a, b)**	**E(a, b)**	**O(a, b)**		**A(b, a)**	**I(b, a)**	**E(b, a)**	**O(b, a)**
		E O	O(a, c)	**A(c, b)**	A(c, a) I		E O	
		O(c, a)		**I(c, b)**	I		O(c, a)	
E O	O(a, c)			**E(c, b)**	O(a, c)	O(a, c)		
O(c, a)				**O(c, b)**				

All tables assume a syllogism relating a and b in the first premise and b and c in the second.

Premises and conclusions are encoded according to the following scheme:

A(x, y) encodes "all x are y". **E(x, y)** encodes "no x are y".
I(x, y) encodes "some x are y". **O(x, y)** encodes "some x are not y".

The letter alone (**A**, **I**, **E** or **O**) is used to indicate **both** possible orderings of the terms.

Johnson-Laird's figures are arranged as:

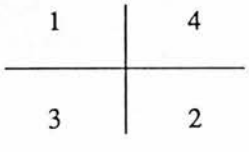


Table 4.1. Valid conclusions for each combination of premises.

The first of these factors relates to the fact that there are five possible relationships between two sets A and B (See Fig. 4.1). Sentences of syllogistic forms give information about which relationship applies in any particular case, but nonetheless, with the exception of the **E** form, remain ambiguous. Thus an **A** premise such as "all whatsits are doobries" is ambiguous with respect to whether there are any doobries that are not whatsits (See Fig. 4.2). The second factor is that even unambiguous interpretations of the premises can usually be combined in several ways. Thus knowing that X and Y are both disjoint from Z does not restrict the relationship between X and Y at all (See Fig. 4.3). These factors combine to give the range of situations in which the premises hold, and both must be taken into account if this range is to be completely evaluated.

Table 4.1 shows valid conclusions for 27 of the 64 possible premise pairs, of which only 19 fit within the traditional restriction on the order of the terms in the conclusion. However, this is not entirely unproblematical. Logicians have no doubt that the premise "some of the whatsits are doobries" is consistent with the premise "all of the whatsits are doobries", and so does not imply that "some of the whatsits are not doobries". However, many people without logical training disagree, or at least are inconsistent or consciously unsure. One could argue that these people simply use a different logic (or speak a different language) that is not in the normal currency of logicians. However, a more commonly accepted solution to this discrepancy between logical and everyday meanings is to invoke the principles of cooperation that Grice (1975) suggested govern all normal use of language. In particular, he suggested that one should make statements that are maximally informative. In the current case, this Gricean rule would imply that, if all the whatsits had been doobries, the premise would have said "all", but since it does not, "some of the whatsits are not doobries" probably also holds.

Wason and Johnson-Laird (1972, pp 143 - 147) report an experiment, conducted by Jean Waddington, which is relevant to this. Subjects were presented with a set of Euler circles illustrating all possible relationships between two sets, and asked to indicate which relationships were consistent with specified sentences. Waddington found that subjects tended to interpret particular premises as inconsistent with the corresponding universal premise. However, even though subjects were told to ignore the actual words used to present the sentences, Waddington found that when the generalisation of the particular sentence was a sensible statement about the world, this tendency was over-ridden and the sentence interpreted the way the material suggested. Thus while few subjects would indicate that the diagram showing set inclusion was consistent with "some books are novels", most would recognise it as consistent with the (logically equivalent) "some beasts are animals". Here, the content of the proposition, which by definition is not among its logical properties, affects its interpretation, a result that tends to count against the "different logic" account. In any case, the resolution of this debate has no direct empirical consequences since although the approaches differ in the logical status they assign to them, they both predict that subjects will accept a range of conclusions that logicians generally regard as invalid.

The problem regarding the quantifier "some" is not the only area of uncertainty in the logical interpretation of the categorial syllogism. In the case of the **1EA** syllogism, with the premises

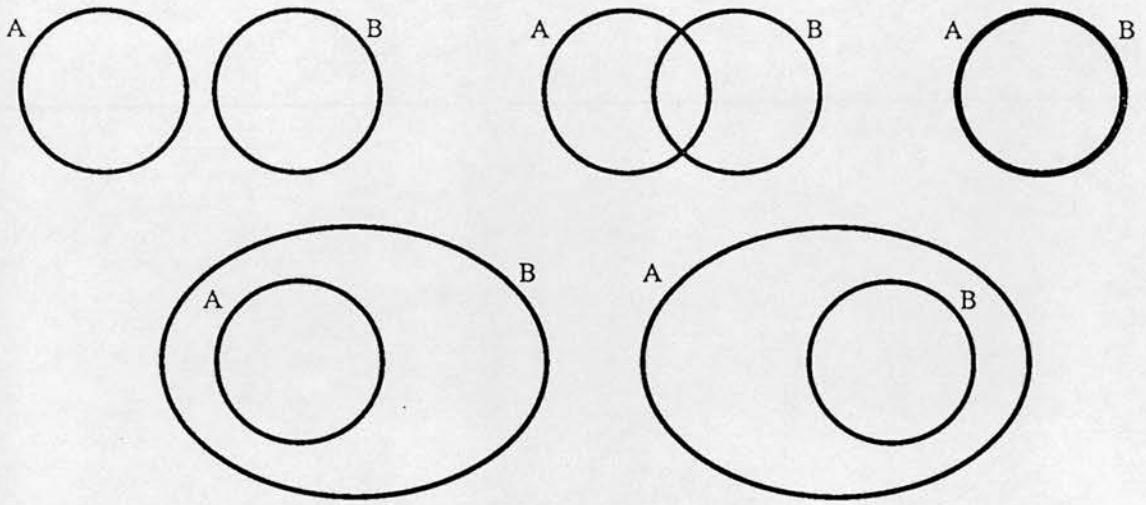


Fig. 4.1: The Five Possible Relationships Between Two Sets

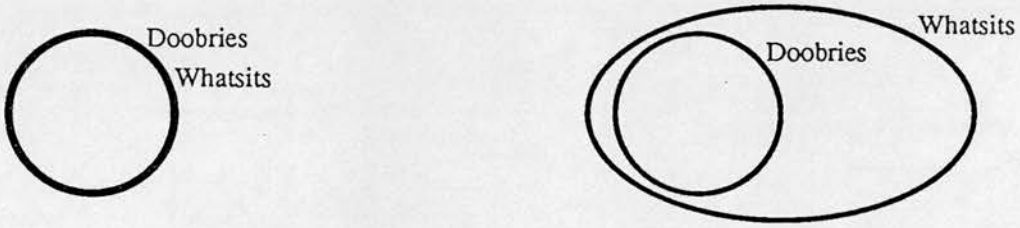


Fig. 4.2: The Ambiguity of "All Doobries are Whatsits"

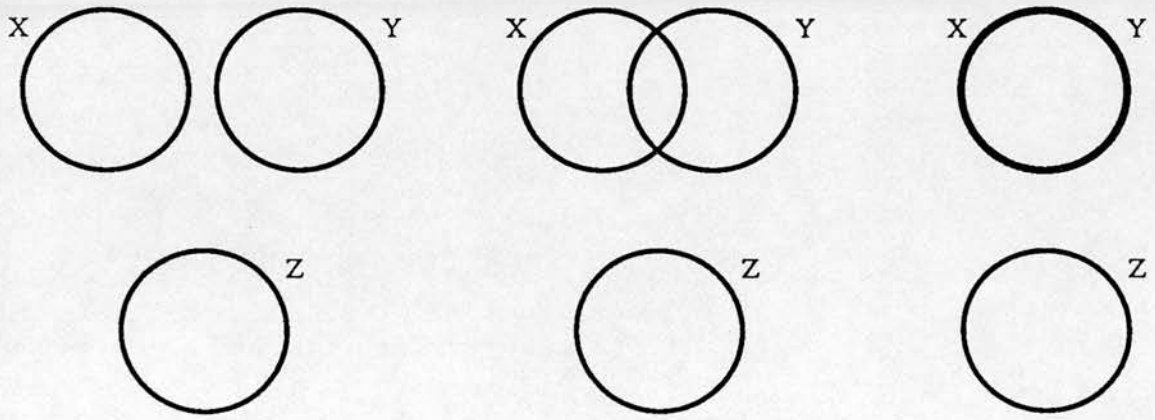


Fig. 4.3: The Uncertainty Arising from Premise Combination

None of the whatsits are doobries
All of the doobries are thingies

the conclusion

Some of the thingies are not whatsits

is deemed valid. Most people accept this, especially when it is accompanied by a justification along the lines of "specifically, it is those thingies that are doobries that are not whatsits". However, to see this conclusion as valid, we have to assume the traditional (Aristotelian) interpretation and read "all of the doobries are thingies" as meaning "there are doobries, and everything that is a doobry is a thingy". Taking a universal (**A**) proposition to imply a particular (**I**) proposition in this way is in line with the traditional logicians *square of opposition* (Copi, 1968), but is not without problems. Russell, while describing the formal defects of Aristotle's logical system, uses the example of the (Johnson-Laird) **4AA** syllogism to point this out:

> If I were to say "All golden mountains are golden", "All golden mountains are mountains", therefore "Some mountains are golden", my conclusion would be false, though in some sense my premises would be true.
>
> (Russell, 1946, ch XXII)

Modern logic attempts to block this unacceptable deduction not by denying the truth of the premises, but by recognising that an A premise is true when applied to the empty set. Under this interpretation, the premises of our problematical syllogism do not say that there are necessarily any doobries that are thingies, or more perspicuously, that there are any doobries at all. Both the premises are consistent with there being no doobries, although the conclusion is not, and hence the inference is invalid. Similar arguments apply to the **2AE**, **4AE** and **4EA** syllogisms.

There would seem to be three moves open to someone attempting to explain the widespread acceptance of a deduction that modern logic regards as dubious. The first would be to suggest that the Aristotelian reading is correct, and that modern logic is wrong to say that the universal quantifier can be applied to the empty set, although Russell's example of the golden mountain is a very strong counter-example. Alternatively, one can once again invoke the Gricean rules, and maintain that the existential assumption ("there are some doobries") is justified by any unhedged mention of the predicate: "if there were not any doobries, they would not have been mentioned". Finally, one might maintain that the presence of the words "of the", which Russell omits, makes a significant difference to the meaning of the premise, although it could be argued that this really only strengthens the Gricean effect. Without taking a stand on the logically correct reading, there is no empirical evidence to suggest that subjects without logical training are not effectively utilising an Aristotelian approach, and this is certainly tacitly accepted in all psychological work on syllogisms.

4.3. Why Do We Study Syllogisms?

Learning how subjects solve syllogistic reasoning problems under laboratory conditions is clearly not an end in itself. Furthermore, since **explicit** formal reasoning is not a common form of human behaviour, the explanation of it hardly warrants the amount of attention that syllogistic reasoning has received. It should be clear, then, that the syllogism is being used as a specific domain to either suggest or support psychological theories relating to mechanisms that have wider applicability. Unfortunately, theorists rarely make clear precisely what they think those

mechanisms are.

One possibility is that syllogisms exercise skills in abstract problem solving. The syllogism is to be seen as a formal problem, and the behaviour observed from an untrained subject reflects the mechanisms that are available to that subject for the solution of unfamiliar, abstract problems. In this light, the syllogism has the advantage of being a task that subjects readily grasp and of offering, as Johnson-Laird points out, a range of difficulty able to exercise most subjects at the limit of their ability. What is more, the methodical relation between the figures and moods of the different syllogisms readily provides a means for the individuation of the factors contributing to the behaviour.

This certainly seems to be close to the approach of the earliest workers. Speaking of her logic students in college, Wilkins says "To many of them to determine the exact meaning of a statement seems to be a thoroughly new task" (Wilkins 1928, P5), while Woodworth and Sells are less definite, saying

> True though it certainly is that the syllogism is a tool of logical analysis rather than a diagram of any typical reasoning process, the use of syllogisms as problems appears to be a promising lead in the experimental study of thinking.
>
> (Woodworth and Sells, 1935. P451)

which is perhaps best read by assuming that the "thinking" involved is conscious, deliberate abstract problem solving. More recently, Erickson (1974) thought it necessary to allow his subjects to either draw Venn diagrams or refer to lists of formal rules, and Newell (1981) uses the task to illustrate his ideas on problem-solving.

However, this view is not universal and many theorists seem to think they are exercising and probing mechanisms closely related to those involved in everyday living. They do not dispute that (externalised) verbal reasoning is rare, but suggest that (subconscious) **internal** reasoning employs a similar mechanism. This reasoning, typically using implicit premises, is seen as playing a crucial role in the functioning of everyday thinking. Thus Johnson-Laird offers the following example:

> Imagine the following scenario:
>
> Person A asks: Where's the university?
> Person B replies: Some of those people are from there.
> Person A goes up to the group of people indicated by person B and asks them the same question.
>
> A's behaviour depends on a chain of inferences that includes at its centre the following deduction:
>
> Some of those people are from the university.
> Any person from the university is likely to know where the university is.
> Therefore, some of those people are likely to know where the university is.
>
> The first premise derives from Bs reply; The second premise derives from general knowledge; the conclusion derives from a process of reasoning.
>
> (Johnson-Laird, 1983. P23)

Henle (1962) explicitly goes still further and suggests that everyday behaviour directly depends on a syllogistic deductive mechanism, citing (P373) Aristotle's *practical syllogism*, which he exemplifies as "For example, when you conceive that every man ought to walk and you yourself are a man, you immediately walk". Having quoted this, and similar examples for drinking and making cloaks, Henle goes on to say

> It is difficult to see how the individual could cope with the ordinary tasks of life if the practical syllogism embodied techniques which are not, as one author quoted above [Lefford] puts it "the common property of the unsophisticated subject", if, indeed, it were not a natural mode of functioning of the conscious mind.
>
> (Henle, 1962. P374)

Clearly almost any cognitive activity can be expressed in terms of the combination of premises, since, as Hayes-Roth points out,

> Scientists have known for many years that one can convert any fully described process into a procedure that operates on propositionally structured data. In sum, propositional descriptions per se do not constitute models, they simply reexpress observations in a formal syntax.
>
> (Hayes-Roth, 1979)

However, this does not mean that it is necessarily a helpful way of analysing the situation. Johnson-Laird may well be correct when he says

> Whenever an argument about a specific entity hinges on a general assertion, the chances are that its deductive form is that of a syllogism:
>
> .
> .
> .
>
> 2) Any point on which a player serves out of turn is a "let".
> A player served out of turn on this point.
> Hence, this point is a "let".
>
> (Johnson-Laird, 1983. P71-72)

It may indeed be that, when the justification for a decision is expressed in terms of logical theory, its form is syllogistic. But this has no implications whatsoever about the use of any kind of syllogistic process in the actual execution of the decision. To see this, consider a syllogistic description of a thermostat reasoning thus:

All rooms that are too cold need the heating turned on.
This room is too cold.
Therefore
This room needs the heating turned on.

We could therefore describe the operation of the heating system as a result of a (practical) syllogism. However, it is generally accepted that an explanation in terms of a bimetallic strip is more perspicuous. There is no doubt that it is possible to give an account of syllogistic reasoning in terms of premise combination. However, the premises being combined according to the most parsimonious account, may not be those of the syllogistic problem being considered.

One could argue that the situation is significantly different in the case of human subjects because, unlike thermostats, they are themselves inclined to speak of their internal operations in terms of propositions. However, this is not a compelling argument. Asking someone to explain or justify their actions probably would elicit accounts phrased in terms of beliefs and inferences. But there is nothing to say that the subjects' explanation of their behaviour is necessarily accurate. Evans (1982) argues persuasively for such a *dual process* explanation, with the explanations of problem solving seen as being generated by a separate mechanism from the actual conclusions:

> Wason and Evans (1975) have shown that protocols may reflect a rationalisation of a reasoning response rather than an introspection of a thought process. According to this view, the conclusion is arrived at by a process of which the subject is unaware and which may be non-logical. The subject is, however, highly logical in constructing an explanation consistent with his response.

(Evans, 1982, P80)

Indeed, there is good reason to be suspicious of the logical structure of the subjects' explanations. Most people are unfamiliar with information processing terms, and lack a well thought out architecture of mind to anchor their application to explaining cognitive behaviour. As a result, they simply have no alternative way of expressing the motivations of behaviour.

Evans's argument for the dual process theory is reinforced by the phenomenon of *confabulation* observed in experiments on "split brain" patients (See Section 2.1). Gazzaniga and LeDoux (1978, pp 15, 131-132) report observations of the vocal side of the brain telling, and apparently believing, a story to explain the subject's behaviour, although the story is often complete nonsense. This can be seen as not a new behaviour pattern generated in response to a pathological situation, but as merely the exaggeration of a normal process of our everyday lives. If this is the case, the "logical" form of the processes that people describe need not reflect the way the problem was actually tackled, which might well be unavailable to introspection. Instead, it could arise from confabulation, as subjects attempt to describe their own behaviour with no more access to the real mechanisms than the experimenter. As a result they produce an explanation shaped largely by their ideas of what mechanism they "ought" to have employed. The generally accepted method for describing how to get from a set of premises to a conclusion is as a sequence of deductive steps, and therefore this is the sort of explanation that subjects will produce.

It should be born in mind, however, that while subjects' reports are not necessarily accurate and introspections can of course be disregarded, such a move brings with it an added burden for the theorist. Eventually, someone somewhere must explain not only how the problem was really solved, but also how the subject's beliefs to the contrary arose.

The balance between these two views of syllogistic reasoning – as an exercise in problem solving or as exemplifying natural thought – is never discussed. Theorists currently working in syllogistic reasoning assume that their experiments give them direct access to (something close to) the mechanisms of everyday thought or reasoning. In particular, they believe they reveal something of the processes underlying the way we understand language and go about our daily lives. If, possibly because of the dominance of logic as a means of expressing cognitive activities, one views cognitive processes in terms of the manipulation and combination of premises, this may well seem natural. Syllogisms are, after all, (among) the simplest kinds of combinations of propositions, and therefore commend themselves to study. However, there are many aspects of experiments that involve working with numerous pairs of peculiar nonsense sentences that might be expected to touch on problem solving skills, and no attempt has been made to justify ignoring the influence that this might have.

4.4. What Is Required of a Theory of Syllogistic Reasoning?

There has been considerable research into the mechanisms involved in syllogistic reasoning, which will be reviewed in the next chapter. A number of opposing theories have been advanced, but before considering them it is perhaps worthwhile to review the criteria that should be applied to their evaluation. Clearly many factors must be taken into account, some of which are related to syllogisms, while others are related purely to what it is to be a satisfactory psychological (or just

plain scientific) theory per se. Johnson-Laird (1983, P65 - 66.) suggests that 7 criteria are relevant to the evaluation of a theory within the current paradigm (in the sense of Kuhn, 1970) of cognitive science, and these will serve as a skeleton for the following discussion of the matter.

> 1) A descriptively adequate theory must account for the evaluation of conclusions, the relative difficulty of different inferences, and the systematic errors and biases that occur in drawing spontaneous conclusions.
>
> (Johnson-Laird, 1983. P65)

This is an encapsulation of one of the most obvious requirements of a theory: it should fit the data. Theories that show why all fish have legs are not useful. On the basis of this criterion all theories except Johnson-Laird's are deficient, in that they fail to account for certain "systematic biases" that have been observed in subjects reasoning behaviour.[3]

However, the reason most other theories fail to account for these biases is simply that they were formulated before the biases were discovered! In asking for a fit to the data, it must be remembered that the data are always changing. Yet Johnson-Laird's phrasing of this criterion, explicitly listing the range of phenomena that a theory must address, invites one to overlook this. Inevitably subsequent workers will reveal new phenomena, every bit as worthy of explanation as any observed so far, and about which no existing theory, Johnson-Laird's included, will have anything at all to say.

While the discovery of new phenomena will obviously present difficulties for any theory, there is a closely related but far more subtle problem. The development of the field will often highlight previously unconsidered features of experimental data. For instance, Dickstein (1978b, P542) argues against a theory (see below) because his experiments do not support its predictions about the relative difficulty of types of syllogism. In fact, as originally formulated the criticised theory makes no mention at all of relative difficulty – it simply was not a factor that its proponents said anything about. Johnson-Laird's second criterion can be seen as requiring just this extensibility from other theories. He suggests

> 2) The theory should explain the differences in inferential ability from one individual to another.
>
> (Johnson-Laird, 1983. P65)

This is a recent trend in psychology, and individual differences are simply not addressed by most other theories, which deal solely with the prediction of statistical distributions of responses.

How does recognising that the normal progress of science will inevitably render any theory inadequate affect the way a theory should be evaluated? It seems unreasonable to reject it simply for failing to cover phenomena that were unknown or unconsidered when it was formulated. Nevertheless, this clearly represents a shortcoming, and the theory must be extended in some way. Most often, this is done not by the theory's original proponents, who are frequently no longer active (at least in the particular field), but by its opponents, who extend it with the intention of

---

[3] Specifically, those that constitute the *Figural Effect*, which Johnson-Laird identified. It will be discussed at length in Chapter 6.

showing that (their extension of) it is incapable of dealing with the new data.[4] Almost inevitably there is more than one way that a theory can be extended, and this swiftly leads to the dethronement of a family of pretenders to the title of the old theory.

How can a theorist circumvent this population explosion? Given that it is impossible to predict all subsequent empirical discoveries or shifts in theoretical standpoint, a realistic goal would be to aim for a situation where all (or at least most) workers are agreed on how a theory should be extended. This requirement is clearly similar to the requirement for extensibility raised in the discussion of the motivation for the study of syllogisms. Johnson-Laird's third criterion leans in the same general direction:

> 3) The theory should be extensible in a natural way to related varieties of inference rather than apply solely to a narrow class of deductions.
>
> (ibid)

Johnson-Laird himself appears to have been concentrating on extension to the problems of reasoning with ill-defined quantifiers, such as "most authoritarians are dogmatic", which is clearly vital given his belief that the mechanism of syllogistic reasoning plays a part in everyday thinking. However, it should be clear that the need to fit constantly changing data demands that a theory be capable of changing just to stand still.

Given that extensibility is a necessary goal, how should it affect the construction or evaluation of a theory? Every theory will have some features which the theorist, or those within the relevant paradigm, consider central, and others which are less important. The former are those that give the theory its character, while the others are simply parameters which tailor the fit of the theory to the world. Only if these types of features can be clearly distinguished will the integrity of a theory be preserved when it is extended. To produce an extensible theory, then, a theorist should endeavour to make clear which features can be varied to fit the data, and what ramifications this variation has within the framework of the theory. Within a computational account, this amounts to a clear statement of the cognitive architecture being assumed.

Johnson-Laird's fourth criterion, like his third, concerns matters not dealt with by most previous theorists:

> 4) The theory should explain how children acquire the ability to make valid inferences.
>
> (ibid)

Psychologists currently accept that whenever any mechanism is proposed for the production of behaviour, it is both valid and important to question how it came to be there. However, the explanation of the ontogenetic development of the mechanism, which Johnson-Laird demands, is only half of what is required. It leaves open the option of placing the explanation of everything within the black box provided by inheritance. In short, the requirement that the mechanisms proposed by ontogenically acceptable is only as strong as the complementary demand for phylogenetic acceptability.

---

[4] This is illustrated by the work, described in the next chapter of, Dickstein (1975) and, to an extraordinary degree, Revlis (1975).

To see this, consider a theory based on the suggestion that subjects solve syllogistic problems by recognising the figure and mood of the premises and producing their conclusions by a process akin to looking it up in an innately provided table. In this case, the behaviour would be seen as arising from a "logic organ", in much the same way as Chomsky (1980) argues for a "language organ". Following Chomsky's lead, the observed development of logical abilities in children could be ascribed to the maturation of this logic organ. This theory has one significant weakness: how did the logic organ arise? This is not to question that genetic encoding is adequate for passing on such an organ – there is no reason to suppose that reasoning would be any more complex to encode than a liver or an eye, or that syllogistic reasoning is inherently more complex than dam making in beavers or nest-building in birds. The question concerns how such a mechanism could evolve. The survival value of syllogistic reasoning ability is not obvious. That of the inaccurate reasoning that most people exhibit is even less apparent.

The need to consider the origin of proposed psychological mechanisms is also pointed out by Newell. Discussing the flow-diagrams that typically form the framework of theories of cognition, he stresses

> the need to understand where these flow diagrams come from – not to understand where the theorists get them but to understand where subjects get them.... Can we understand how the task gives rise to the flow diagram?
>
> (Newell, 1981. P696)

However, the way he proposes that we seek this understanding bears closely on the other criteria suggested above – not only the obviously related differences between individuals, but also the extensibility of the theory, both to related phenomena and changing data.

Newell identifies three areas of psychological research as investigating closely related phenomena: Reasoning, Problem Solving and Decision Making. Although this might be expected to give rise to inter-related theories or research programs, he considers that in fact each field is proceeding as if the others did not exist. Without any overall coherence or unity, psychologists in these areas study cognition using "just one damn task after another... without essential orderliness" (ibid, P.694). This has dire consequences since, he suggests, each of these disparate tasks quite independently becomes the subject of a theory, the framework of which is usually encapsulated by a flow-diagram.

> My question is: Where did these flow diagrams originally come from? They are different for every task, though they all surely bear a family resemblance. Theorists seem simply to write a different theory for each task. Details get filled in experimentally, but the frameworks (i.e. the flow diagrams) are just written down.
>
> The diversity of these flow diagrams is connected to the fragmentation of cognitive studies. Diversity per se does not cause the trouble, for the tasks themselves are indeed diverse and the theories must reflect that. The difficulty lies in the emergence of each of the microtheories full blown from the theorist's pen. There is no way to relate them and thus they help ensure the division of the study of human cognition into qualitatively isolated areas.
>
> (ibid, P694-5)

Crucially, Newell's plaint is not that the theories are incoherent or incomplete – careful use of modern computer simulations allows a theorist to demonstrate precisely what behaviour his theory is capable of producing. Nor is it that the sub-processes invoked are ill-defined or incomputable – the procedures invoked within the simulation itself can serve as at least examples of the kind of

processes envisaged. The problem is with the flow of control, as captured in the structure of the charts themselves:

> The diversity of these flow diagrams arises in large part because these diagram theories incorporate the detailed structure of each task within the very fibre of the theory. They are a version of the magicians trick – by the time the theory emerges, the scientific magic has already taken place.
>
> (ibid, P696)

What is needed but signally lacking, Newell claims, is is an over-arching, domain-independent theory, and in its absence psychology can offer only a set of ad hoc stories. There is an obvious similarity between such a cross-domain theory and the idea of a paradigm put forward by Kuhn (1970), though arguably the latter is of even grander scale and wider scope. Newell's central purpose is to highlight the pointlessness of working in isolated domains, and to emphasise the necessity for a unifying theory that can encompass several tasks and can constrain, at least roughly, the general shape of an acceptable theory.[5] In doing so, it would distinguish the crucial features from the empirically determined parameters – the very thing required by the four criteria discussed above. Thus the four criteria discussed so far can be seen as manifestations of the same underlying requirement for an adequate theory to be clear about the paradigm in which it is set.

Whatever paradigm a theorist may be working in must explain, or (more generally) just accept, a number of facts about the way subjects approach syllogistic reasoning experiments. Undergraduates in Western universities performing such tasks are (almost) universally willing and able to reason about abstract situations, such as imaginary rooms full of people. That is, they realise that it is possible to manipulate Xs and Ys without considering what they are, though such realisation is seldom accompanied by aptitude. They also realise that the Xs and Ys in one item have nothing in common with the Xs and Ys in any other, and that they must ignore the fact that their uncle Clive is a chemist who keeps bees when assessing the truth of the sentence "no beekeepers are chemists". Finally, though syllogism experiments do not reveal it, almost everyone can understand the idea of a counter-example defeating a conclusion – that there are circumstances when the exception is not allowed to prove the rule.

Where does this grasp of the fundamental ideas of abstract reasoning come from? The instructions at the start of the experiment surely cannot teach such complex notions, only indicate that they are appropriate, so subjects must already have the relevant concepts. Since studies by Luria in 1931 to 1937, there is considerable evidence to suggest that it arises from the educational process (see Luria, 1977 and Scribner, 1977). In any case, wherever these skills or abilities might be acquired, they contribute to, if not constitute, what might be termed the subject's "rationality". This is an evocative term, which proved to be a tar-baby for early workers in the field (see Section 5.2). It is also nebulous, even ambiguous, and the abilities listed above do not seem to be what Johnson-Laird has in mind when he introduce his fifth criterion:

---

[5] Newell's own proposed unifying, cross-domain theory, involves not just a common cognitive architecture, but also commonality between the procedures employed in different domains. It is centred on the idea of the *Problem Space*, and is discussed in Section 5.4.3, as is his illustration of its applicability to the domain of syllogistic reasoning.

> 5) The theory must allow that people are capable of making valid inferences, that is, they are potentially rational.
>
> (Johnson-Laird, 1983. P65)

Although it may not be possible to specify precisely what rationality is, Johnson-Laird's "that is" is certainly a mischaracterisation – it is not by any means synonymous with being "capable of making valid inferences". Picking responses to syllogistic reasoning problems by pulling the handle on a fruit machine and suitably interpreting the winning line would often produce a valid inference, but nobody would want to say that this was a rational process. And, crucially, it would remain an irrational process even if it seemed accurate – even if, over a large number of trials this method had never been observed to suggest an incorrect response. Suggesting that "all artists are beekeepers" is a valid conclusion because there is a melon on the middle reel of a gambling machine is simply not a rational thing to do.

In contrast, however, consider the situation where the fruit machine is in fact (known to be) not a simple mechanical gambling engine, but a terminal to a sophisticated computer, able to understand syllogistic premises spoken in its presence and programmed to stop its reels in such a way as to indicate the correct conclusion. Given this situation, offering conclusions on the strength of the position of the reels becomes a perfectly rational thing to do. Admittedly not **reasoning**, but undeniably **rational**. These examples show that whatever rationality is, it is undeniably tied up with making inferences for acceptable **reasons**.

What counts as an acceptable reason for doing something? The ideal answer would surely refer to a justified belief that the method would lead to a satisfactory result. But such a belief can only come about as a result of the application of the method to other tasks – one cannot have faith in the outcome of an untried procedure (unless it arose by applying trusted approaches to reliable information about a task). Thus rationality can be seen to be in essence the tackling of problems by the application of methods that have already suggested their value, their general applicability. Actions are rationalised (justified) by showing how they were the result of applying (specialisations of) generally acceptable methods. This also illustrates how arguments can be disproved by the use of a counter-example to demonstrate the lack of generality of the methods involved.

There is still another feature of the concept of rationality. Consider the situation of a person offering conclusions to syllogistic reasoning problems on the basis of the roll of a die (and, naturally enough in a philosophical discussion, giving correct answers). This is not obviously a candidate for rational behaviour. However, this can be changed by the addition of a belief that the die is not in fact a passive piece of wood, but is actually a small robot, once again able to interpret syllogistic problems and able to twist itself in flight to ensure that it lands indicating the correct response. Or, if preferred, a belief that the die is under the control of the spirits of the dead. It should be clear that the rationality of the "reasoner" depends in a complex way on whether it is s/he or the observer who holds the cited belief, and also, in the former case, on the observer's attitude towards it.

This suggests that an observer views behaviour as rational to the extent that the reasons for the behaviour are acceptable, or intelligible. Rationality is an intentional property, and granting or denying, or even considering, the rationality of a system is only meaningful to the extent to which

the system is seen as having motives or reasons. Notice that, within the view of the worl espoused by Dennett (1978), there is no requirement to show that the system "really" has reasons, and certainly no need for it to be able to express them. It need only behave as though it had them. Once this is recognised, the futility of discussing the logicality or rationality of a mechanism is revealed. To the extent to which something is regarded as mechanistic, it is regarded as deterministic, and thus incapable of exhibiting behaviour indicative of intentionality. A mechanism cannot be rational since it cannot do things for the right reasons, simply because it does not do things for any **reason** at all. Its behaviour can certainly be described as following the laws of physics, but there is no ground room for debating rationality there – causes are not reasons. A broken clock is just that: broken. There are no irrational thermostats.

Regardless of its precise definition, people are undeniably "rational", at least in the sense that they can recognise the existence, and abundance, of reasoning errors. As a result, a great deal of effort has been devoted to the formalisation of the reasoning process, which has given rise to formal logics, Venn Diagrams et al. It therefore seems reasonable to require of a theory, as does Johnson-Laird, that:

> 6) The theory should shed some light on why formal logic was invented and how it was developed.
>
> (ibid)

The thrust of this criterion is in the second clause – "how it was developed". It is difficult to see how any theory can fail to motivate the development of some kind of formal logic. It need only capture the fact that reasoning is difficult, and ascribe the rise of logic as the response of ingenious people trying to ease a heavy burden. More telling is to ask how the underlying processes have **shaped** the formal logics that have arisen, and what the style of modern logics tells about the "natural" approach to thinking.

Johnson-Laird's final criterion, while undoubtably a worthy objective, is hardly a suitable criterion by which to judge a theory.

> 7) The theory should ideally have practical implications to the teaching of reasoning skills.
>
> (ibid)

Clearly a good theory of a group of phenomena will lead to greater understanding of the area, and may then allow workers to go on and reach otherwise difficult and important goals. But this facilitation is only a side-effect of a theory and cannot be demanded of it, since to do so prejudges the relevance of the theory to the goal, and thus delimits its area of applicability. Relativity overthrew the classical laws of conservation of matter and conservation of energy because it was judged to be a superior theory. But undeniably **absent** from the criteria by which this judgement was made were its practical implications for the destruction of whole cities with a single bomb.

4.5. The Categorisation of Errors

One of the necessary features of any theory is that it must define, possibly implicitly, what phenomena it is attempting to describe, what factors it will take into account and what idealisations it takes itself to be making. This will simultaneously define what factors are not being taken into consideration, and thus will effectively be adding "noise" to the predictions of the theory by

causing the true behaviour to deviate from the predictions of the theory. For example, when considering an arrangement of strings and pulleys, the physicist may well ignore cable masses and deformation, friction and angular momentum in the pulleys themselves and air resistance to motion. S/he will often arrive at predictions for the behaviour of the system which are nearly right; for instance predicting that a weight will fall to the ground in ten seconds when in fact it takes twelve. However, there are circumstances where the ignored factors become so significant that the predictions from the idealisations made in the theory become grossly inaccurate; for instance where the idealisation predicts that the weight will fall, but the friction in the actual system is enough to prevent it moving at all.

In these cases the physicist is left with two choices. One possibility is that s/he can "explain" the observed behaviour by ascribing the deviation from the theory to some external factor(s) about which the theory is not able to say more because they are beyond its terms of reference. Alternatively s/he may attempt to expand the theory to take account of what s/he believes the (important/relevant) perturbing factors to be, and thus produce a new and more powerful, more general, theory. Clearly the latter of these choices is greatly preferable, although equally clearly it is not always practicable, since the inclusion of the relevant factors may be an extremely difficult task indeed.

Nowhere is this more obvious than in the field of Psychology. Here, the most widespread methodology is to assume that the general abilities of an individual mind can be thought of as arising from the coordinated interplay of different mechanisms, and that progress can be made by analysing these component abilities independently, much as the kidneys can be studied independently from the heart. This means that a typical psychology experiment is geared to revealing something about a small but (the psychologist believes) relatively independent ability (or skill or mental organ) of the subject. The problem is that the experimenter can do nothing to prevent other mental processes within the subject from interfering with the experiment, since he has no influence over the normal functioning of the subject's brain. The best that can be achieved is to minimise the opportunities for interference by designing the experiment so as to avoid providing (or at least control for) stimuli for procedures, responses or abilities other than the one under study. Thus many experiments measure subjects' abilities to recognise particular sounds in sequences of meaningless words, perceive depth in random patterns of dots or memorise lists of nonsense syllables.

Unfortunately, when this assumption of independence seems to be incorrect, the psychologist is often unable to pursue the ideal course of generalising his theories – the mechanisms underlying the neglected factors are of impenetrable complexity. As a result, psychologists tolerate unexplained separation between the predictions of a theory and the observed behaviour of subjects. This distinction – which is eloquently supported by, among others, Chomsky (1965) – is justified and encapsulated in the distinction between the *competence* and the *performance* of a subject. The former in some way represents the true ability of the subject, which it is the aim of the psychologist to explain. In actual behaviour it is only exhibited when beset by unimportant, inexplicable, performance factors. However, this is a very coarse categorisation, and there is a real

gain from a finer analysis. When analysing a subject's performance on a task, four distinct categories of failures can be identified, each with its own predicted distribution of errors.

(1) *Competence* errors: The task requires the subject to undertake something of which he, she or it is fundamentally incapable. Goldfish will not record very high scores in syllable repetition experiments.

(2) *Execution* errors: Occasions where, for reasons that are unexplainable by the theory in hand, the behaviour observed deviates from that predicted. The attack of hiccups during speech illustrates this kind of error within linguistics, since, presumably, the probability of suffering a spasm of the diaphragm does not depend on the syntactic complexity of the sentence being uttered. This is not to say that they are uncaused, but that the causes are believed to be factors outside the domain of current interest. This means that, when the deviations from predicted behaviour they cause are analysed within the terms of reference of the theory, they will appear to have a truly random distribution, with occurrences statistically related only to time.

Execution errors are obviously undesirable, not only because they represent places where the theory's predictions are incorrect. To the extent that a theory suffers from execution errors, that theory is incomplete, in that it is neglecting factors that are relevant to the behaviour it seeks to explain. Clearly only the perfect theory of life, the Universe and everything will be completely free from outside influences. All theories put forward in the foreseeable future will be bound to suffer execution errors to some extent. Nonetheless, despite their inevitability, their presence still gives a measure of the completeness of a theory and the independence of the range of phenomena and factors it has chosen to address.

(3) *Capacity* errors: Occasions where the psychological theory breaks down, yet neither at random nor because of any qualitative feature of the phenomena involved. Instead, some (possibly fundamental) quantitative limit has been reached or exceeded. Everybody is able to reliably repeat a sequence of five random syllables, so in some sense people have the competence to hold and recall syllable sequences. However, nobody could manage the same feat for a sequence of twenty five syllables, despite the fact that the problem is clearly the same in almost all respects. There is a clear sense in which the universal failure on the longer sequence is not just the effect of random, extra-theoretical factors causing "performance errors", but instead reflects a different **kind** of error – something closely akin to a limitation on the subject's competence.

(4) *Ability* error: Occasions where subjects fail to achieve the performance of which theory suggests they are capable, simply because they have not learnt some necessary skill or piece of knowledge. In some sense every Englishman has the competence to understand a sentence in Chinese. The reason he cannot is hardly outwith linguistic theory, nor a reflection of any kind of memory limitation or whatever. It is simply that he doesn't know how.

The execution / capacity error distinction is obscured by the fact that the approach to a resource limitation often results not in a sharp breakdown as the relevant limit is exceeded, but in inconsistent performance, and an increase in errors that cannot be accounted for by any feature

within the theory (i.e. an increase in execution errors). In short, people pushing themselves to their limits often make mistakes. Thus subjects manage to recall some strings of twelve syllables, but not others of only eleven. This kind of behaviour can readily be explained by suggesting that the resource limit available to any given process is continually fluctuating under the influence of extra-theoretical factors, and that more of the resource that is required, the more likely it is that one of the fluctuations will mean that the limit is exceeded. However, this blurred limitation does not give the theorist a complete carte blanche. Unlike the case of execution errors, capacity errors should monotonically increase as the limit is approached. This means that a theory should be able to at least predict where capacity errors are likely to occur, and their failure to occur where, and only where, predicted is undeniably a fault in the theory.

Finally, notice that both kinds of errors predict statistical noise in addition to a monotonic increase with a particular feature. In the case of capacity errors, it is the quantity that is limited, in the case of execution errors it is time (i.e. opportunity). If there is a statistically significant concentration of performance errors that is not in line with any such trend, it reveals a factor that is inducing errors, but is not explained by the current theory – a relevant factor that is currently being overlooked.

4.6. Summary

The categorial syllogism is one of the simplest deductive forms possible. As such, logicians made it the focus of their attempts to analyse reasoning, and the formalised interpretations of its sentence forms became central to the developing discipline of logic. Psychologists began studying how people tackled syllogisms over 70 years ago. Initially, they seem to have regarded the activity as an example of problem-solving, though more recently, and possibly as a result of the use of formal logic techniques to model mental cognitive processes, it has been seen more as simply an exercise of everyday reasoning. The results of their labours are reviewed in the next chapter. The theories presented there must be evaluated in line with a number of criteria. Clearly matters of empirical accuracy cannot be overlooked. However, at least as important are the inter-related matters of explaining individual differences and the acquisition of the skills involved. Both of these are closely tied to the extensibility of the theory, which in turn depends on its relation to some other, domain-independent theory or paradigm. Finally, the explanation of subjects incorrect reasoning must account for their poor performance while still allowing that the subject can recognise correct reasoning and achieve expertise.

CHAPTER 5

Theories of Syllogistic Reasoning

5.1. Early Work: Wilkins, The Atmosphere Effect and Illicit Conversion

The oldest referenced psychological work on the categorial syllogism was performed in Germany by G. Storring between 1908 and 1925. (references 2 - 7 in Woodworth and Sells, 1935). However, it seems to have had negligible influence on subsequent work, since Woodworth and Sells are the only authors to acknowledge its existence, and although they claim to have studied Storring's results, they had no acknowledged influence on their paper.

The first work to achieve significant recognition was **Wilkins (1928)**. Wilkins work was motivated by a desire to investigate the factors that impaired the performance in formal logic classes of some otherwise intelligent students. She believed that they had difficulty with the abstraction of the logical content of the problems from their presented form. Thus she carried out an experiment that involved presenting logically identical problems to subjects in four different forms:

(1) Using familiar terms, carefully chosen to minimise the extent to which subjects would be influenced by their knowledge of the truth or falsity of the statements (e.g. ''All the pink cups are Mary's'')

(2) Using symbolic material (e.g. ''some a's are b's'')

(3) Using unfamiliar terms: words which were, or looked like, technical terms outside the subjects' vocabulary (e.g. ''some boarddentates are otitic'')

(4) Using familiar material, chosen so that the subjects knowledge of the material would tend to mislead them about the validity of the deduction (e.g. ''No oranges are apples, No lemons are oranges, therefore no lemons are apples'')

Wilkins employed twenty logical problems, mainly categorial syllogisms, but also including transitive inferences and deductions from single premises, which she used in two (similar) forms, giving two related sets of problems. Each part of her experiment entailed presenting subjects with both sets of problems using one of the four types of material – eight batches of twenty items in all. Wilkins(1928) contains the complete listing of the actual test items used.

Eighty college students took part in the experiment, each attempting both forms using two types of material at each of two sessions. Each set of 20 items was presented in a booklet, with a time limit of twenty minutes. Each test item consisted of a pair of premises (or in some cases just one statement) and three possible conclusions, and subjects had to mark each of the three to indicate whether or not they thought it was the result of a valid deduction from the

premise(s) – i.e. the experiment asked subjects to **validate specified conclusions.**

Wilkins main result is well known. She states

> Ability to do formal syllogistic reasoning is much affected by a change in the material reasoned about. The easiest material is the familiar and concrete. The most difficult is the unfamiliar (long words). The symbolic material is almost as difficult as the unfamiliar. The suggestive material is more difficult than the familiar but not so difficult as the symbolic and the unfamiliar.
>
> (Wilkins, 1928. P77)

This finding is in line with other explorations of the effect of material on reasoning performance, such as the Wason Four-Card Test (Wason, 1966, 1977), and also served to reinforce everyday observation. For this reason, subsequent experimenters have been careful to minimise the potential influence of the meaning of the premises they present their subjects. Wilkins also found that

> The standings of individuals in tests of syllogistic reasoning are affected to some extent by changing the material reasoned about. The correlations between one part of the test and another are positive and high, but not so high as between different forms of the same part. The standings of certain individuals are very little affected by the change in the material, whereas the standings of others are changed very materially.
>
> (Wilkins, 1928. P77)

In other words, different individuals are affected to differing extents by a change of material.

The rest of Wilkins's results are related to the categorising of subjects reasoning errors in terms of the *fallacies* of traditional logic, and in the correlation of syllogistic reasoning performance with intelligence tests (which she finds is "marked, but not high", though for symbolic material the correlation is "decidedly higher than with any other kind of material used" (ibid))

Although Wilkins thoroughly investigated the factors influencing errors in syllogistic reasoning, she conspicuously did not offer any theory to account for them. The first such theory to be put forward was the Atmosphere Effect, first proposed in **(Woodworth and Sells, 1935)** and slightly modified in Sells (1936) and based on the experiments reported therein. Sells used symbolic material to present 300 syllogisms, including conclusions, 71 of which were valid. Ninety "educated adults" were asked to indicate whether the conclusion was "'true' or 'false'". They commented that the order of the premises seems to have little effect upon the conclusions that subjects will accept. The following year, Sells (1936) reported a similar experiment, using 169 syllogisms, of which 127 were invalid. He reported his results in terms of the 16 possible premise mood combinations, ignoring the effect of figure. He found that **I** conclusions were always accepted more readily than **A**, and **O** almost always accepted more than **E**. Further, the most popular incorrect response was always **I** or **O**.

The theory Woodworth and Sells advanced was essentially aimed at explaining subjects' errors on syllogistic reasoning tasks, which they ascribe to three factors:

(1) The ambiguity of "some". As was mentioned above, logicians always interpret "some" to mean "at least some", while people without logical training often interpret it to mean "some, but not all".

(2) The principle of "caution", whereby subjects prefer to offer particular, as opposed to universal, and negative, as opposed to positive, conclusions.

(3) The "atmosphere" of the problem.

However, both the flavour and the majority of the predictive power of the theory come from the atmosphere component.

The basis of the atmosphere effect is that each premise of a syllogism has an *atmosphere*, which can be either positive or negative, and either universal or particular (e.g. an **A** premise has a positive universal atmosphere). When two premises are combined in a syllogism, it in turn acquires an atmosphere based on its premises, according to "secondary hypotheses". Despite the fact that Begg and Denny (1969) deny it, and are often referenced as providing a more succinct account of the mechanism, Woodworth and Sells are prefectly explicit about how the atmospheres of premises are combined:

> In general formulation the secondary hypotheses suggested for application to syllogistic reasoning are (1) that a particular premise creates a "some" atmosphere, even though the other premise be universal, and (2) that a negative premise creates a negative atmosphere, even though the other premise be affirmative.
>
> (Woodworth and Sells, 1935. P453)

When applied to Sells data, this rule alone serves to predict the most accepted conclusion to almost all the premise combinations. The exceptions arise because in some cases (i.e. **AA, AE, EA** and **EE**) the Atmosphere Effect predicts a universal conclusion, whereas Sells always found particular conclusions more popular. This discrepancy was probably the motivation for the principle of "caution" mentioned in (2) above.

Chapman and Chapman (1959) subsequently suggested that the exceptionless preference for particular conclusions could be an artifact of the experimental design. They argued that it arose when a particular conclusion was offered for validation with a premise pair that supported a universal conclusion. Subjects would be counted as supporting the particular conclusion even though they knew perfectly well that a universal conclusion is appropriate. However, Woodworth and Sells were aware of this, since they write:

> There is however excellent logical ground for such acceptances, once "atmosphere" is recognised; for an **A** conclusion logically implies its subordinate **I**, an **E** its subordinate **O**. Whenever a subject would have accepted an **A** conclusion, he should be willing to accept an **I** also.... Thus nearly all the acceptances not referable to mere atmosphere can be explained by atmosphere plus a little logic, or by atmosphere weakened by caution, or by the ambiguity of "some".
>
> (Woodworth and Sells, 1935, P458)

Notice that their "atmosphere plus a little logic" explains the acceptance of a particular conclusion via the acceptance of a universal conclusion, and thus requires the process of validation to proceed via the solution of the problem. This is the only hint at any kind of process that might be involved in the solution.

Finally, it is worth stressing that Woodworth and Sells explicitly state that the Atmosphere effect is not the only mechanism acting when subjects work with syllogistic material. When initially introducing the effect they say that "when a subject does not see the relationships clearly, he is influenced by the atmosphere" (P453), and they go on to comment on the differences between the acceptance rates of a given conclusion when it is valid and invalid, saying that

> The difference, Valid - Invalid ... can be regarded as a measure of the amount of acceptance of the valid conclusions which is due to seeing the logical relationship. That is, while this group of college students showed a 96 percent acceptance of the valid **A** conclusions from **AA** premises, they also showed a 36 percent acceptance of invalid **A** conclusions from the same type of premises, because of atmosphere or other illogical factors, and therefore they can be credited only with a 96 - 36 = 60 percent net score for genuine understanding of the relationships in the **AAA** syllogisms.
>
> (ibid, P459)

Thus, even though the Atmosphere Effect predicts the majority of valid conclusions, Woodworth and Sells explicitly accept that it acts together with some other (in some sense "logical") mechanism, about which they say nothing further. However, as Wason and Johnson-Laird (1972, P134) comment, "this cannot be said with such certainty of their successors", and in many ways there are two theories of Atmosphere. Woodworth and Sells spoke for Dr. Jekyll, who leads confused reasoners when the problem is beyond them, but most other theorists see only the wicked Mr. Hyde, a fiend who destroys their ability to reason at all. Thus, by 1971 when Ceraso and Provitera find evidence that the Atmosphere Effect is supported by some other factor, it is a discovery worth reporting, and they write:

> Note that the number of Ss who give atmosphere responses is considerably greater in those cases where it is also the correct response than where it is not. In the latter case note that the correct response is often chosen. This would argue that at best only some of the responses are a result of atmosphere, and that there is a fair amount of valid reasoning occurring.
>
> (Ceraso and Provitera, 1971. P404)

Evans (1982, P82) too presents the Atmosphere Effect as a complete solution mechanism and suggests that "the irrational conception of man" offered by the theory was one of the main factors motivating Chapman and Chapman to challenge the theory after more than twenty years of widespread acceptance.

In opposing the "irrational" Atmosphere theory, **Chapman and Chapman (1959)** begin by suggesting that the tendency towards particular conclusions that Sells had observed was an artifact of his methodology.[1] To avoid this problem, they performed their own experiments, asking their subjects to choose from a list of options comprising the four (classically acceptable) possible conclusions and "none of these". Since they were concerned with reasoning errors, they concentrated their attention on those syllogisms that have no valid conclusion. Their results were very similar to those of Sells, and differed in two main ways :-

(1) They do not replicate Sells's observed consistent bias towards particular conclusions even when only universal premises are involved. This is in line with the suggestion that the effect was indeed an artifact of Sells's methodology.

(2) On three syllogisms they differ markedly from Sells's results. For the **EE** syllogisms, Chapman and Chapman found that **E** was the most common response, while for the **EO** and **OE** syllogisms, **E** and **O** were equally common. Sells, in contrast, found only **O** responses to all these forms.

---

[1] Although as was shown above, Woodworth and Sells fairly explicitly admit this.

Chapman and Chapman place much emphasis on this second difference between their results and Sells's, and on the strength of these 3 differences (out of the 14 premise pairs that they examine) they claim that

> Since the atmosphere predictions are not substantiated, we must look for other principles of explanation.
>
> (Chapman and Chapman, 1959. P224)

They then go on to propose an alternative account of reasoning on syllogistic tasks which captures all their results. The most emphasised aspect of their alternative theory is the suggestion that errors arise from subjects' misinterpretation of the premises. Specifically, subjects are inclined to accept the converse of A and O premises, which are not valid i.e. they *convert* the premises when it is not justified. The traditional name for this error, by which the Chapman and Chapman theory is usually known, is *illicit conversion.*

In support of illicit conversion, Chapman and Chapman call upon the Wilkins's (1928) data, and point out that the acceptance of the converse of an A premise was a common error by her subjects (items A14 and A*6 in Wilkins, 1928). However, they fail to point out that, although this conversion was a common error, it was still exhibited by only 25% (i.e. a minority) of Wilkins's subjects, whereas over 80% of Chapman and Chapman's subjects fall victim to the errors that they wish to ascribe to illicit conversion. Their case is weakened still further by the findings of Jean Waddington. Her experiment (see Section 4.1.2) showed that although one quarter of her subjects accepted the conversion of O premises as valid, only one in eight accepted the conversion of an A premise.

Wason and Johnson-Laird (1972, P149-150) have argued that conversion is more likely within the context of a syllogism than when the premise is interpreted alone. They describe (ibid, P55-59) experiments in which they examine the propensity of subjects to convert sentences of the form "If A then B" to get "If B then A", which is invalid (consider "if there is a rainbow, then it is daytime"). They find that

> When a seemingly difficult or insoluble problem can be solved by assuming that a conditional does imply its converse, then subjects are prepared to do so.
>
> (ibid, P59)

This leads them to suggest that since illicit conversion often turns an invalid syllogism into a valid one, it too should be encouraged by the desire to solve the problems. This theorising is significantly unclear when it comes to the interpretation of the mechanism by which conversion is encouraged. Wason and Johnson-Laird mention the factor of "cognitive load", which might suggest that illicit conversion was some kind of capacity error. However, for the most part their discussion is phrased in terms of subjects adopting it as a strategy which enables an otherwise (for them) insoluble problem to be solved.

Although illicit conversion is the emphasised feature of the theory, it accounts for the most common errors for only six of the fourteen types of premise combinations that were considered, and in doing so is in complete agreement with the Atmosphere Effect. The remaining results, including all those that were said to damn the Atmosphere Effect, have to be explained by a second mechanism, *probabilistic reasoning.* Here "probabilistic" refers not to whether or not a particular inference will be made, but to the making of an inference that is not definitely, but only probably,

true. Chapman and Chapman suggest that everyday reasoning will condition people to accept such inferences, saying

> Such thinking is not unreasonable; rather it is the reasoning process by which most science progresses. Thus a chemist might reason as follows: "Yellow and powdery material has often been sulphur. Some of these test tubes have yellow powdery material. Therefore some of these test tubes contain sulphur." This is an invalid **III** syllogism of the second [traditional] figure, yet the conclusion has some probability.
>
> (Chapman and Chapman, 1959. P225)

Having presented this example, Chapman and Chapman go on to suggest that a similar style of reasoning, coupled with illicit conversion, will account for the errors they observed. In particular, this kind of reasoning is only seen as happening when the premises are in the 3rd figure,[2] where the middle term appears in the second (predicate) position in both premises. Where they are not presented like this, they must be appropriately converted, possibly illicitly. Finally, they say there is an ambiguity in the process of probabilistic inference, such that when **O** and **E** premises are combined, the result can be either an **O** or an **E**. It is this feature that explains the bimodal distribution of responses that the Atmosphere Effect cannot.

The two theories proposed as a result of this early work, and above all the attempt to prove the superiority of Illicit Conversion over the Atmosphere Effect, dominated work in syllogistic reasoning for more than forty years. The Atmosphere Effect highlights a reliable pattern in most common reasoning errors, and cleanly captures their dependence on the surface features of the premises – i.e. whether either is a negative or a particular statement. In contrast, Illicit Conversion attempts, with limited success, to tie the abundance of unsound reasoning to the (mal-)functioning of some kind of logical process – specifically, a tendency to inappropriately convert the premises of the problem. Unfortunately, in order to explain the observed range of responses, this central mechanism had to be supported with others of tenuously defended "logicality". However, this was done in the firm belief that to do otherwise would be to deny man the ability to act for rational reasons. Since then, the need to establish this ability has been a prime motivation for many workers, a matter which will be discussed further in the next section.

5.2. Henle and Logicality

The next reported experiments on syllogistic reasoning are described in **Henle (1962)**. However, this work is unlike other work within the paradigm, in that rather than attempting to gather data on the **results** of subjects' reasoning, Henle attempts to focus directly on the processes by which they are produced. She does this by presenting subjects with examples of syllogistic arguments and asking them to evaluate their validity and explain their justification for their conclusion. She then examines the subject's justifications in the hope of revealing the processes involved in the production of the conclusion. This method is justified only by the far from obvious assumption that the subjects not only have access to the processes by which they reason, but also are able and willing to report them.

---

[2] According to Johnson-Laird. It is the 2nd figure in their (traditional) notation

Without offering any justification, Henle departs from the normal experimental practice of all other workers in the field, before or since. She makes no attempt to minimise the influence of subject's prior knowledge and prejudices by presenting problems using symbols or "content free" familiar terms. Instead she presented her subjects with narratives which included a deduction about matters of everyday experience, and asked them to comment on the validity of the deduction. She says her "instructions included an explicit statement that the logical adequacy of the arguments was to be judged, not the the truth of the statements" (Henle, 1962. P369). However, Wilkins (1928), which Henle notably does not reference, has demonstrated that such instructions do not remove the influence of the material.

Finally, Henle's presentation of her results is unlike that employed in any other experiment in the field. She provides no data as to the accuracy or distribution of her subjects' judgements. Instead, she offers only a selection of "illustrative data" in the form of extracts from subjects' justifications of their conclusions which she uses to support her categorisation of the reasons for subjects' errors. She motivates this by saying

> As many authors have shown, the incidence of error in deductive reasoning depends on the form of the syllogism and its contents, as well as on instructions to the subjects. Quantitative results would have relevance only to the particular conditions studied here, whereas an enquiry into the nature of the error obtained might be of more general interest
> (Henle, 1962. P370)

It is indeed true that the results of syllogistic reasoning experiments may be influenced by many factors apart from those that the experimenter is manipulating. The same is true of experiments of all sorts in all sciences. However, to deduce from this that the actual results are irrelevant outside the particular experiment, and thus that only the particular experimenter's interpretation of them is significant, is to thwart one of the most fundamental processes of scientific research. It is only by comparing the results achieved in replications and variations within a paradigm that it is possible to isolate the reliable effects that science is attempting to identify from those that are an artifact of the details of the experimenters' methods and possible errors.

By choosing a paradigm which deliberately induces factors that most experimenters have attempted to exclude (i.e. meaningful material), and then presenting only sketchy details of both her materials and results, Henle has minimised the usefulness of her experiments to the point where they make no contribution to mainstream syllogistic reasoning research.

Henle's principle objective was to establish the "rationality" of subjects' reasoning in syllogism experiments, and in particular to reject the "irrational" atmosphere effect in favour of Illicit Conversion. She begins with an extended presentation of the two possible views of the relationship between logic and psychology. The older view is that logic is "the science of thought", and that "the rules of logic are those of human understanding", whereas the "modern" view is "the assumption that logical principles are irrelevant, if not antithetical, to much actual reasoning". She summarises the decision to be made between these viewpoints thus:

> Two clearly contrasting alternatives thus present themselves: Is logic (or Aristotelian logic) largely irrelevant to the thinking process, or is it concerned with the laws of thinking? Since we will here be concerned only with deductive reasoning, we may reformulate the question more specifically in these terms. But since, as has so often been pointed out, the premises from which we reason are commonly not spelled out, since our

> inferences so frequently appear as enthymemes**, this fact must be taken into consideration. We may ask: If we know the premises – tacit as well as explicit – from which a person reasons, can we put the process in syllogistic form? Do the rules of the syllogism describe processes that the mind follows in deductive reasoning even when the syllogistic form is not explicitly employed?
>
> ** a syllogism that is incompletely stated, in which one of the premises of the conclusion is tacitly present but not expressed, is called an enthymeme. (Cohen and Nagel, 1934, P78)
>
> (Henle, 1962. P368)

What can it mean to say that the process a subject employs in reasoning is "described by the rules of the syllogism"? If we are dealing with an explicit argument, then we can say whether or not each stage is admissible according to the rules of our agreed syllogistic logic. We must be able to find a rule of inference to warrant the deduction of each theorem from previously accepted theorems. Thus drawing of an **I** conclusion from two **I** premises, as might be predicted by the Atmosphere Effect, would be seen as an unacceptable, illogical, step. If we make assumptions about the progress of an internalised deduction, there would seem to be a natural extension of this approach that will provide the distinction between theories of mental activity that Henle is concerned about. This is the course that Henle adopts.

Using "illustrative examples" of the results of her idiosyncratic experiment, she suggests that subjects are, in fact, reasoning logically when they perform syllogistic tasks, and that their poor performance is due to one of four other factors, namely

(1) Failure to accept the logical task.

> More specifically, this means failure to distinguish between a conclusion that is logically valid and one that is factually correct or one with which the subject agrees
>
> (Henle, 1962. P370)

(2) Restatement of a premise or conclusion so that the intended meaning is changed. For instance a particular premise might be treated as a universal.

(3) Omission of a premise. "Occasionally a subject employed only one of the presented premises" (ibid).

(4) Slipping in of an additional premise. Henle says this was infrequent in her experiment, but more common in everyday discussion. "Premises may be added that are so commonplace, so much taken for granted, that they escape attention" (ibid, P372).

These four categories of error are not well defined, since all instances of (1) or (2) can be seen as instances of (4), while (3) alone would seldom permit a conclusion to be drawn from the single remaining premise. In any case, the core of Henle's approach is that subjects making errors should be seen as reasoning correctly, but not from the premises presented in the problem. Thus she concludes that:

> If we consider the materials and task as they were actually understood by individual subjects, we fail to find evidence of faulty reasoning. It must be concluded that the presence of error does not constitute evidence that the laws of logic are irrelevant to actual thinking. The data tend, rather, to support the older conception that these laws are widely discernible in the thinking process.
>
> (ibid, P373)

However, this approach will not do. Achieving a logically justifiable sequence of inferences is only part of the behaviour that is needed to account for rationality. To see this, consider the following inference and its justification:

Some typists are women.
Some secretaries are not women.
therefore
All secretaries are women.

Provided one is prepared to ignore the second premise (by 3), then the common sense piece of knowledge that "all secretaries can type" can be combined (because of 4) with "all typists are women", a slightly misinterpreted form of the first premise (by 2), to validly deduce the postulated conclusion. Thus although the actual conclusion may be invalid, the above argument outlines how accepting it could, for Henle, be seen as "rational".

Such coverage makes it hard to see how the use of the term can be seen as meaningful at all. Nonetheless, Henle sees the justification of such a chain of inferences as central to the defence of rationality. Evans (1982) attacks this approach, pointing out that in any reasoning situation one can either determine the axioms employed or the rules of deductions used **but never both.** However, this is precisely what Henle tries to do, and the attempt is futile. Henle aims to buy a form of rationality for her image of mankind, but the price is common sense.

This is not the only flaw in Henle's arguments. She argues that Illicit Conversion is in some sense more logical than the mechanisms that theorists assume give rise to the Atmosphere Effect. Certainly Chapman and Chapman themselves try to present their theory as in some sense a "logical" alternative to it. Before giving examples of probabilistic reasoning, they introduce the idea, saying:

> It is known that conclusions are often reached by probabilistic inferences ... and our Ss had no way of knowing that all but strict deductive reasoning is disallowed in the syllogistic game.
>
> (Chapman and Chapman, 1959. P224)

Recall that subjects were being confronted with test items expressed in symbolic form. This suggests that their previously acquired knowledge and prejudices about the subject material (typically groups of consecutive letters of the alphabet) is unlikely to exert much influence on their reasoning. Therefore, this remark is actually saying that the subjects were not reasoning according to the rules of standard logic, which would have lead them to valid conclusions. Instead, they were combining the premises according to some other set of rules which they considered applicable to reasoning situations. Clearly, these rules warrant the combination of premises to produce conclusions that logicians do not support as valid. Chapman and Chapman propose just such a set of rules (i.e. those of probabilistic inference) which they defend by arguing that although they are not valid, they are often correct, implying that although it might not be "logical", it was at least "reasonable".

However, the Atmosphere Effect also specifies such a set of rules. Moreover, no appeal to evidence from everyday life can distinguish the "logical" Illicit Conversion from the "illogical" Atmosphere Effect, since almost all valid inferences (as might be experienced in everyday life) support both. Since this support makes both proposals equally reasonable, any difference in their

logicality must be justified, as suggested in Section 4.4, from the motivations for taking them. There are two possible motivations for taking illogical steps towards a rational conclusion, neither of which offers any credence to an attempt to distinguish the two proposals on the grounds of "rationality". The first is to assume that the subject knowingly chooses to make these unjustified deductions. This amounts to saying that the subject, by opting for a "likely" deduction, demonstrates a complete failure to grasp the concept of logical validity, although this is surely a central feature of rationality. Alternatively, one can assume that the subjects simply employ these strategies unavoidably, either because they know no better or because they are somehow "natural" or "built-in", which means that the subject is unavoidably, fundamentally, incapable of sound reasoning. In any case, there is no principled distinction between the two causes of the bad deductions – the subject simply does not follow logically defensible deduction rules, and this is illogical behaviour. The difference between the theories that has attracted the most attention, especially with respect to the the rationality of conceptions of man, simply amounts to the fact that Woodworth and Sells do not describe their rules as a logic.

5.3. Distinguishing the Theories

Distinguishing these two theories – Illicit Conversion and the Atmosphere Effect – became a central focus for the psychology of syllogisms. On the basis of the criteria presented in Section 4.4, there is little to choose between them. In terms of predictive accuracy, Chapman and Chapman rejected the Atmosphere Effect on the grounds that it could not predict the support for **E** conclusions that they observed in their experiments. However, this is not a sound argument, since to precisely the extent this prediction failure is a valid argument against Woodworth and Sells, one can argue against Chapman and Chapman that **their** theory cannot predict the equally significant fact that Sells's subjects **didn't** accept the **E** conclusions. Note that the difference between their experimental method (selecting the correct conclusion from a range) and Sells's (marking conclusions as valid or not) does not explain this difference. Although Sells's solution marking would lead to more support for **O** conclusions **in addition to E** conclusions, there is no obvious way it should inhibit the accepting of **E** conclusions that illicit conversion predicts. In other words, the experimental support for the theories is precisely equivocal.

How, then, do they compare according to the yardstick offered by the other criteria? As far as the explanation of rationality, both theories present a picture of a mechanism that can give the right answers, but which has a set of errors to which it is prone. Neither theory makes any mention of how the mechanism is acquired nor how or whether it differs between individuals. Considering extensibility, neither theory really seeks to clarify its relation to its paradigm. Woodworth and Sells liken the errors produced to certain linguistic slips and thus imply that the errors arise from giving too much weight to the consideration of local or surface features of the task (the "atmosphere" of the premises). In contrast, Chapman and Chapman suggest no related errors, although their discussion of the motivation of the invalid reasoning procedures – they often give correct results – invites the generalisation that subjects will often adopt unjustified but effective procedures. To the extent that this means anything at all, it seems to throw confusion on how it is that subjects can be said to be rational. In conclusion, on the evidence presented so far the

argument between the theories was equivocal.

5.3.1. Response Distributions

The most direct attempt to break this symmetry was an attempt to investigate the differences in empirical findings to which Chapman and Chapman attached such importance. Ignoring Simpson and Johnson's opinion (quoted below) that error distribution alone was inadequate to distinguish the theories, **Begg and Denny (1969)** attempted a replication of Chapman and Chapman's experiment, although they actually presented all 64 possible premise pairs and ignored those with valid conclusions. Again subjects had to mark which of the four (traditionally) possible conclusions they thought was valid, though since there was no "none of these" option, they could only indicate that they thought that no conclusion was appropriate by putting a question mark in the margin.

Begg and Denny's results were very similar to both Chapman and Chapman's data and Sells's. Calculating product-moment correlations between the sets of data for each premise form pair indicated two areas of disagreement between the three sets of data.

(1) Where the premises were both universal, Begg and Denny found, like Chapman and Chapman, that universal conclusions were more popular than particular conclusions. This is in contrast to Sells's result, and once again suggests that this is an artifact of his experimental method.

(2) For the three results (**IE, EO** and **OE**) where Chapman and Chapman had differed from Sells, and on which difference they had claimed empirical evidence against the atmosphere effect (see above), Begg and Denny's results agreed with those of Sells.

From these results they conclude that

> The atmosphere hypothesis still functions as an effective predictor of behaviour, both in the present study and in that of Chapman and Chapman. As emphasised earlier, this does not constitute support for either the hypothesised mode of reasoning expressed in the atmosphere position or the illogical conversion process. This must await research aimed at uncovering the processes involved in reasoning, rather than assessing the end product.
>
> (Begg and Denny, 1969, P354. Anglicised spelling)

Another attempt to distinguish these theories in terms of their ability to predict distributions of responses comes from **Revlis (1975)**. He suggests that the difficulty in distinguishing between the accounts was largely attributable to the uncertainty in their formulation, which meant that

> research has failed to compare them on the critical conditions and in their intended senses. With these considerations in mind, the present paper undertakes to recast the two hypotheses as processing models of syllogistic reasoning and to provide a first [sic] test of them.
>
> (Revlis, 1975. P181)

This is, of course, an example of the phenomenon, highlighted in Section 4.4, of a theory being extended by someone other than its original proponent, and indeed, it can be seen as an object lesson in how unsatisfactory this practice can be.

Revlis begins by outlining the Atmosphere Effect, describing it, as indeed did its original proponents, as a mechanism that comes into play when the reasoner is in difficulty. However, he then offers the "strong criticism" that it

> neither embodies a model of reasoning nor explains the presence of errors in syllogistic reasoning. That is, the hypothesis does not specify why reasoners should fail to grasp the nature of the relationship between subject and predicate terms, nor does it provide a statement of the overall probability of an error for any specific problem.
>
> (ibid, P182)

Having highlighted these weaknesses, Revlis then proceeds to extend Woodworth and Sells's theory by making it into a processing account with the "advantage" of being "sufficiently detailed to make predictions concerning the solutions to every syllogism" (ibid, P183). The predictions are simply the result of assuming that subjects generate their response to every premise pair by selecting its features – universality and negation – on the basis of those in the premises. Revlis also predicts the relative likelihood of subjects deviating from these predicted conclusions, at least by making execution errors occurring in proportion to the amount of processing. In doing so he is making use of the fact that his algorithm for the feature selection process requires more processing steps to take the default value of a feature when there is a mismatch between the premises than it needs to accept a value shared by both, although no justification for this disparity is offered.

In the same way, Revlis goes on to present the theory of Illicit Conversion, which is similarly criticised (P185) for its failure to specify the conditions under which conversion will occur. He then proceeds to offer a "processing" version of this theory too, opting for the "strongest interpretation" of the possible conditions for conversion: namely, that the subject will convert every sentence read – i.e. both premises and (in a selection experiment) candidate conclusions. If this does not allow a conclusion to be selected or given, Revlis postulates that the subject will, if there is time available, re-read the premises and this time reason with them the way they are actually presented. If this still does not yield a conclusion, or if there was insufficient time for the second reading, the subject will simply guess – a feature Revlis defends on the grounds that subjects will be reluctant to not get an answer to a test question. As a result, this theory distinguishes three categories of syllogism: those that change the conclusions they support if the premises are converted, those that do not, and those that do not support a conclusion under any circumstances. Revlis terms these "DIFFERENTS", "SAMES" and "NONES" respectively.

Having presented his "recast" versions of these two accounts, Revlis goes on to describe an experiment he conducted in an attempt to distinguish them. There were 64 test items, each a premise pair containing symbolic material (i.e. letters of the alphabet). These were based on 17 forms that support a conclusion with terms in the traditionally allowable order and 18 that do not, with repetition being used to provide a total of 32 of each sort of item. This material was presented to two groups of 25 students, who were in both cases required to select either one of the four possible conclusions forms or "none of the above is proven". One group were required to select their response within 15 seconds, the other within 30 seconds, although Revlis found no significant difference between them on any measure.

On the basis of the results of this experiment, which he does not present, Revlis compares the performance of his two extended theories. The feature selection (Atmosphere) approach predicts 71% of subjects' total responses, although this falls if the predicted conclusion is not valid. There is also evidence that its accuracy is impaired if the premises differ in their features, in line with his notion that resolving such discrepancies requires effort. The conversion model, however, is only

61% accurate overall, although this reflects the averaging of 88% accuracy for "DIFFERENTS" premise pairs that support a conclusion only when converted – and 45% for those that are unaffected by the process. The two accounts differ in their predictions on nine premise pairs.[3] For five of these they make mutually exclusive predictions, with 89% of subjects actually giving the response predicted by feature selection, making it "overwhelmingly the better predictor of the reasoner's behaviour" (Revlis, P191). On the other four, the conversion model makes two predictions, one of which is different from the result of feature selection. In these cases the two conversion predictions between them cover marginally more responses that the single feature-based response – 85% versus 79.4%.

Revlis's conclusions from this data are astonishing. He notes that the conversion model predicts 10% fewer responses overall, and in contrast to the feature model is

> incapable of predicting the precise response that a reasoner will make on five of the problems which critically distinguish between the conversion model and the feature selection model. However, the conversion model should not be dismissed solely on this basis. While the feature selection model is more accurate in its predictions, its motivating principle appears to be contradicted by the data: If subjects are making their decisions based on feature-matching and not on logical inference, the model's accuracy should be uniform across conditions... Yet the model is most accurate when its prescribed decision corresponds to the logically correct one... In addition, the model is incorrect in predicting uniform error rates across invalid syllogisms because reasoners make logically correct decisions for NONES [i.e. they (correctly) state that there is no conclusion]. Together these facts suggest that the feature selection model fails to account for the observed rationality of the subjects and that its higher predictive accuracy may be spurious.... While the [conversion] model's predictive accuracy is lower than the feature model, the data do support the motivating principle of the conversion model – namely, that the reasoner's decisions are guided by rational processes.
>
> (Revlis, 1975. pp 191-192)

Finally, Revlis proposes a "combined" model: this is essentially his conversion model, but with the tendency to guess at random when no conclusion can be found being replaced by a selection guided by feature selection. He claims that this combined model predicts 71.8% of subjects responses, although he does not mention that this is precisely the same as the feature selection model alone.

Revlis claims that

> In contrast with the traditional forms of the atmosphere and conversion hypotheses, the present specifications of the feature selection and conversion models make strong, testable, and differentiating predictions concerning performance.
>
> (Revlis, 1975. P187)

This is indeed the case, but they do so only on the basis of Revlis's extensions to them, which he makes no attempt to justify. In particular, he explicitly admits that his interpretation of the conversion theory means that

> conversion is intrinsic to the representation of the premises on all abstract syllogisms, so that the reasoner never makes deductions from the same propositions that are provided him [sic] by the experimenter.
>
> (ibid, P186)

Nonetheless, this is the account he defends on the basis of its "motivating principle... that the

[3] Eight of these have valid conclusions, which means that these comparisons are being made on solely on the basis of Revlis's own (peculiar) interpretation of the original accounts.

reasoner's decisions are guided by rational processes'' although he does so without elaborating on what it is about these processes that earns them this label.

The problem with Revlis's attempt to distinguish between the two accounts is that it is fundamentally misconceived. His final rejection of feature selection is based on the qualitative features of the model it presents, and the fact that the reasoner is seen as taking no account of the meaning of the quantifiers involved. Such meta-theoretical grounds can indeed form the basis of an argument for rejecting a theory, but this is not the argument that Revlis develops. Instead, he presents an experiment which is intended to discriminate between the competing accounts, and which ''overwhelmingly'' supports the feature selection model. Admittedly there are some aspects of the results that feature selection cannot encompass – in particular, that subjects sometimes state that nothing can be concluded for a particular premise pair, or that its accuracy is better where the conclusion it predicts is valid. However, Woodworth and Sells knew, and explicitly stated, that the effect they observed operated in conjunction with some other aspect of subjects' reasoning processes. Indeed, they even suggested that the change in performance with the validity of the conclusion was a measure of logical comprehension! As a result, the Atmosphere Effect was never intended to have anything to say about recognising that there is no conclusion to be drawn, and the fact that it cannot is purely a reflection on the way Revlis attempted to make it into a ''processing account''. It was proposed as a strategy adopted by the floundering reasoner and its accuracy on Revlis's experiment has done nothing to impugn this.

Revlis may well be right when he suggests that subjects are reluctant to say that they cannot draw a conclusion. He has certainly identified a class of problems – those he terms ''DIFFERENTS'' – which do appear to generate significantly different errors from other forms. Regrettably, these ideas are buried among the rhetoric of logicality.

5.3.2. Effects of Training

A more subtle attempt to distinguish between the two theories was made by **Simpson and Johnson (1966)**. They noted that

> Considerable overlap can be seen in the frequencies reported by Sells and by Chapman and Chapman for comparable syllogisms, in spite of the difference in format. There is considerable overlap also in the errors predicted by the two hypotheses. Thus this method [comparing error rates between forms] probably has little more to contribute to the evaluation of the alternative hypotheses. The situation now calls for experimental manipulations from which different effects on errors can be predicted.
>
> (Simpson and Johnson, 1966. P197)

This led them to an experiment to analyse the effects of training subjects to avoid the specific errors that the theories predict.

Simpson and Johnson used ten syllogisms of each of three sorts: Those with valid conclusions, those for which only the Atmosphere theory predicted errors, and those for which only illicit conversion predicted errors. These were presented to five groups of subjects. One group was warned about the Atmosphere Effect and given a chance to practice avoiding it. A second group was told about Illicit Conversion and shown examples of it. A third group received only a general warning that syllogisms were tricky, while a fourth received that and some practice. Finally, a fifth, control, group received no warning at all. Unfortunately, both groups of error-inducing

materials used in this experiment are highly problematical.

When describing their material designed to induce illicit conversion errors, Simpson and Johnson claim, without further explanation, that

> Although many syllogisms give opportunities for invalid conversion, there is only one, AO, that can disclose this error uncontaminated by atmosphere error. It can be written in several variations, however.
>
> (Simpson and Johnson, 1966. P198)

In fact, the "several variations" amount to three. Even if the lexicalisation is varied, it seems likely that ten items that are essentially just the same three problems will induce some kind of learning (at least in the sense of experience-related alteration of behaviour).

Unfortunately, repetition is not the only problem with this group of data. For an **AO** syllogism, illicit conversion predicts an **O** conclusion, but so too does Woodworth and Sells's theory. Hence it is not clear in what sense Simpson and Johnson can see the **AO** syllogism as "uncontaminated by atmosphere error". One possible motivation for this enigmatic statement is suggested by the fact that before they describe the Atmosphere Effect, they say that Woodworth and Sells

> added two supplementary principles to account for discrepant results, but these were contradicted by the results of Chapman and Chapman and may therefore be eliminated
>
> (ibid, 197)

The obvious interpretation is that this refers to the principle of "caution", the motivation for which seems mainly to have been an artifact of Sells's experimental method, and thus indeed absent from Chapman and Chapman's data. But it seems that they are actually referring to the rules for the dominance of dissimilar atmospheres, for they entirely omit this central feature of Woodworth and Sells's theory when they go on to describe it:

> Thus the atmosphere effect can now be rewritten in specific language: when both premises contain the same qualifier, "are" or "are not", or the same quantifier, "all" or "some", many subjects will accept a conclusion which also contains the common term.
>
> (ibid)

Now if this central feature of the theory is omitted, "atmosphere" does indeed predict no error for the **AO** syllogism, but it is not easy to see what feature of the Chapman and Chapman data warrant such an omission.

The problem with the "atmosphere only" group is equally fundamental. Although illicit conversion itself does not predict error responses to the problems used in this group, it constitutes only one half of Chapman and Chapman's theory, and no attention at all is paid to Probabilistic Reasoning.

Simpson and Johnson seem to have completely overlooked important factors in both the theories, and as a result of these serious omissions, their experiment cannot be seen as distinguishing between the two theories, but only as probing the details of either. Nevertheless, there are several points to note about their results:

(1) Correlation of performance indicates that "susceptibility to one type of error is practically independent of susceptibility to the other" (ibid. P198).

(2) The warning about the atmosphere effect was both effective and specific. The group given it was by far the best on the "atmosphere error" section of the test, though no better than the

unwarned control group on the "conversion error" section.

(3) However, warning about conversion errors actually improved performance **less** than the general warnings, regardless of material.

(4) Finally, and perhaps surprisingly, any kind of warning at all **impaired** subjects ability to find the valid conclusions for those problems which had them.

These results led Simpson and Johnson to conclude that their subjects' poor syllogistic reasoning preformance is a manifestation of two independent types of error, and that although their training against conversion errors had not been particularly effective.

> In general, differential training gave differential results. Neither the atmosphere error alone not the conversion error alone can account for these findings.
>
> (ibid, P200)

Unfortunately, the second part of this conclusion is only warranted by their impoverished version of the theories. What they have undeniably shown is that training can remove some errors, which they characterise as "atmosphere errors", without improving performance on others. This suggests that those errors are generated by a distinct mechanism. However, Chapman and Chapman already explicitly say this, the second mechanism being precisely that part of their theory that Simpson and Johnson ignore, namely "probabilistic inference". While this actually differs from the Atmosphere Effect in certain of its predictions (see above) Simpson and Johnson neither take note of this nor report their results so that these distinctions can be analysed.

Dickstein (1975) criticises Simpson and Johnson not only for omitting probabilistic reasoning (see above) but also for the limited range of syllogisms they employed, for including practice items only in the anti-atmosphere training and for assessing the performance of the two groups on sets of syllogisms that differed in difficulty. However, he supported their opinion that the Atmosphere and Conversion theories are too similar in their predictions to be distinguished solely on the grounds of error distributions. Thus he too ran an experiment to compare the preformances of groups of subjects warned against one or other of the two types of error, though he also examined the result of presenting the premises in the reverse order. In addition to the effects of training, Dickstein attempted to distinguish the theories on the grounds of the predictions that they would make regarding the relative difficulty of different types of syllogism, and to do this he had to make assumptions about the processes involved in the theories.

In the case of the Atmosphere Effect, he identified four groups – those where the premises agreed in both aspects of atmosphere (A1), those where they agreed in universality (A2), those that agreed only in the polarity (A3) and those that agreed in neither (A4). He then suggests that the atmosphere effect should be stronger where the premises agree completely, weaker where they only overlap, and weakest of all where they differ on both factors. This leads him to predict that atmosphere errors will decrease from A1 to A4.[4] Similarly, for the Illicit Conversion theory, he distinguished the problems that yielded conclusions after illicit conversion of a premise (C1), and

[4] Note that this prediction was not a part of the Atmosphere Effect as proposed by Woodworth and Sells, who made no mention of how the effect arose. Dickstein's extension of their theory, therefore, illustrates the phenomenon mentioned in Section 4.4: the extension of a theory in line with current ideas, often (as in this case) by its opponent, and with the intention of showing that the extended theory is inadequate.

those that require probabilistic reasoning (C2). He then suggests that there ought to be more errors on C1, since conclusions generated using probabilistic reasoning are only possible, not necessary. This argument is obscure, since it is not clear how the uncertainty of a conclusion can affect subjects told to seek necessary conclusions.

In Dickstein's experiment, all 64 syllogisms were presented, with subjects responding by selecting one of a list of possible answers that included the traditionally possible conclusions and the option that no conclusion was valid. All the subjects were given instructions about what was required of them, but one third were also warned about atmosphere errors, and another third about Illicit Conversion and probabilistic reasoning. Half of each group received problems with the premise order reversed, though the order in which items were presented does not seem to have been controlled or randomised.

Neither premise order nor instructions had any significant effect on valid syllogisms. For invalid syllogisms, however, the group warned about conversion errors performed best while the group warned about atmosphere errors scarcely improved over the control group. This is the complete reverse of Simpson and Johnson's results. Dickstein explains part of this discrepancy by pointing out that his instructions were effective in reducing probabilistic reasoning errors, but errors due to actual illicit conversion were not affected. Since Simpson and Johnson ignored probabilistic reasoning, this suggests a possible reason why they failed to obtain any performance improvement in their "Illicit Conversion" group. However, Dickstein offers no opinion on why his warning against atmosphere errors was ineffectual while Simpson and Johnson's was clearly effective. Presumably this can be put down to the fact that they gave their subjects practice items whereas Dickstein did not.

The order of the premises had an effect only on the "atmosphere" group, which accepted fewer invalid conclusions when the premises were presented in the reverse order.

When it comes to the relative difficulty of problems, Dickstein reports that his results, and indeed those of Chapman and Chapman (1959) and Roberge (1970), show that errors are indeed more frequent on C1 problems than on C2, while what little trend can be perceived among the A categories in the same studies indicates that A4 errors are more common than A1. These trends, he claims, clearly support the Illicit Conversion theory and count against the Atmosphere Effect. However, this is not particularly damaging for the Atmosphere Effect, since the prediction of the trend in error rate is based on the assumption that one particular (though admittedly fairly natural) kind of mechanism is responsible for the effect. Thus its supporters are free to claim that this is a wrong assumption, and Dickstein is wrong to predict trends in difficulty.

The fact that experience (in the form of instruction or practice) can improve performance on certain problems without affecting others suggests that there is more than one mechanism or factor involved in syllogistic reasoning errors. Because of their apparently poor grasp of the theories involved, Simpson and Johnson's experiment did not allow any actual comparison between the main theories. Dickstein's results, however, serve to support the Chapman and Chapman distinction between illicit conversion and probabilistic reasoning, and also indicate that the former is much less influenced by instruction.

5.3.3. Modified Problems

Although Henle's own experiments are of dubious worth (see above), her theories suggesting that subjects reasoning errors arose primarily from misinterpretation of the premises motivated **Ceraso and Provitera (1971)** to seek experimental support for this idea. They worked with syllogisms phrased in terms of the properties of wooden blocks, and presented them both in the normal manner and in a modified form, re-phrased to remove the vagueness of the traditional premise forms. Thus for instance, a traditional **A** premise, such as "all red blocks are triangular", does not determine whether there there are any triangular blocks that are some colour other than red, and they would have been disambiguated in one of two ways:

"Whenever I have a red block it is triangular, and the triangular blocks are red"[5]

or "Whenever I have a red block it is triangular, but there are some triangular blocks that are not red"

They then argue that

> If we can show subjects are reasoning correctly in the modified syllogism case, it would be gratuitous to assume that they were not doing so in the traditional case
>
> (Ceraso and Provitera, 1971. P402)

The experiment employed two groups of undergraduate subjects. One received an (apparently unmotivated) selection of 13 syllogisms in modified form (M) while the other received the corresponding traditional syllogism (T). However, these 13 modified problems were elaborations of only eight traditional problems, which meant that the "traditional" group received several problems more than once. Subjects seem to have been tested individually, with the problems being presented verbally, in a random order for each subject. No information is given as to whether subjects could also read the premises, or could ask to hear them again or similar. They gave their conclusion by marking a pre-printed answer sheet. This gave them the choice of three of the four conclusion forms (with the terms in, apparently, only the order that traditional syllogisms do not allow!) or "can't say". The **O** form ("some A are not B") plays no part in the experiment at all.

Their use of a pre-printed answer sheet with randomised item order raises questions. It is possible that the conclusion to every syllogism related exactly the same terms (e.g. always quantifying how many red blocks have holes), so the array of possible answers was identical in each case, which would lead one to expect very strong interference between items. Alternatively the subjects had to hop about the answer sheet as the items were presented, in which case they could see (and be influenced by) their earlier answers. Neither alternative is entirely satisfactory, and insufficient detail is given to reveal what was actually done.

The least pleasing aspect of Ceraso and Provitera's results is that when identical problems were presented (the same traditional forms corresponding to different disambiguations – see above) distinctly different results were obtained. There are cases where the popularity of solutions to a particular problem changes by about 25% of the total group size. Note that since the presentation order is randomised, this cannot be put down to learning during the experiment, but must reflect

[5] Notice that great pains were taken to avoid the word "all", to the extent that this example is unnatural, arguably to the point of introducing ambiguity.

some real difference in the actual problem items. No mention is made of this matter at all.

A possible explanation for these large variations in responses to apparently identical problems lies in another, unfortunate, feature of their experimental procedure, not mentioned above. Although the description of the experimental method is vague on this point, it seems that subjects were not only read the premises, but were also shown actual blocks that illustrated them. Thus

> For example, the T group might be shown a red block with a hole in it, and be told "all red blocks have holes". They might then be shown a triangular shaped block with a hole in it, and be told, "all blocks with holes are triangular". The task would be to state the relation between red blocks and triangular [ones].
>
> (Ceraso and Provitera, 1971. P402)

No mention is made of what sort of block is shown to accompany a negative premise, whether each premise was always accompanied by the same block, or whether the same blocks were used in both the traditional and modified presentations of the problems.

Ceraso and Provitera make no attempt to justify or explain the use of real blocks in any way. The only remotely plausible motivation seems to be the belief that the sight of such a block would make the task more concrete for their subjects. This objective would be in line with Henle's attempt to phrase her syllogisms in "natural" terms. However, it is hard to accept that showing undergraduate students a specific wooden block would significantly alter either their understanding of the nature of the situation or their grasp of the logical structure of the problem. It might well, however, influence their behaviour by providing a counter-example to a potential conclusion they consider. For instance, where the premise pair is "all red blocks have holes, all triangular blocks have holes", a subject's tendency to suggest that "all red blocks are triangular" will clearly be influenced by whether or not either premise is illustrated with a triangular red block. In effect the physical blocks shown can be thought of as supplying extra premises (e.g. some red block – this one – is triangular), and yet no information is provided to enable one to judge what they might have been or how (or whether) they were controlled. This is a serious deficiency, and it seems possible that this factor accounts for the response variations in otherwise identical problems.

Ceraso and Provitera's main result was that their subjects did considerably better at solving the modified form of syllogism – i.e. that performance improved when the premises were made unambiguous. This allows them to argue that since subjects are not led astray by an atmosphere effect for the modified problems, there is no reason to assume that they are in the traditional ones either. They go on to point out that (for the range of syllogisms they have chosen) the erroneous responses that subjects gave to the traditional problems were often the most common response to the modified forms. They then suggest that this is because subjects are dealing with the traditional premises as though they had actually been presented with the modified premises. They do not recognise the range of interpretations of the traditional premises and thus fail to consider the full range of possible situations that they permit.

There is another important feature of Ceraso and Provitera's results. They point out that when subjects deal with modified syllogisms, their error rates correlate closely with the number of ways that the (completely specific) premises could be combined. This is in line with the observation that the syllogistic reasoning task involves checking the truth of a conclusion throughout a range of possible situations. Modifying the premises removes one of the factors that generate this range,

and thus reduce the search space. In short, subjects do better because the problem is easier! Ceraso and Provitera recognise this possibility as distinct from Illicit Conversion. However, they reject the idea on the grounds that subjects make systematic errors, though they give no further explanation. This suggests that they disregard completely the idea that a defective solution process could be methodical, the result of ability errors. However, this is to be expected given a belief (suggested by using wooden blocks to make the situation "real") that they are exercising the mechanisms of subjects' everyday thinking, and that those mechanisms must allow an essentially rational picture of man.

5.4. Other Theories

5.4.1. Associationism

All of the work mentioned so far was directed at analysing the effects of the form of the premises. Citing no allegiance to any of it, **Frase (1968)** describes experiments on the effect of the figure of the syllogisms, with the intention of demonstrating the applicability to syllogistic reasoning of the ideas of mediated response commonly found in associationist learning theory.

Frase links the presentation of each premise with the learning of an association. Associationist learning theory, developed as a result of experiments on the learning of nonsense syllable pairs, predicts that the two associations resulting from the premises will produce an association between the end terms. The strength of this will depend on the directions of the premise associations, which will be determined by the syllogistic figure of the problem. This association will facilitate subjects' drawing (and validating) of conclusions. In particular, mediation will have most effect in figure 2 (N.B. Johnson-Laird's nomenclature), and least in figure 1, with figures 3 and 4 being intermediate :-).

Frase (1968) presents the results of 2 similar experiments, both framed within the structure of traditional syllogism, with the subject matter being the eleven syllogisms that have the same validity across all figures (of these, one has a valid conclusion). He presented each of these in each figure in a randomised order and interspersed with 11 valid "filler" items. Every conclusion related X to Y, with the middle term always Z. Subjects were asked to decide on the validity of the presented syllogism, and rate their confidence in their answers on a five point scale.

In the first experiment of Frase (1968), two groups of subjects were used, one of which had received some elementary training in syllogistic reasoning, while the other had received only comparable but irrelevant training. A "computer based teaching machine system" was used to present the materials and note the subjects' answers and response times. The results showed that the trained subjects were faster that the untrained group, averaging 209 seconds for 11 syllogisms, as opposed to 284. They were also more accurate, with an average of 1.2 errors compared to 2.6. Frase comments that the response times did not indicate any "warm-up" effects. There was a definite effect of figure:[6] For both groups, figure 2 (I) was most accurately done and accorded the

[6] N.B. The terms in the conclusions offered for validation were always in the same (classically allowed) order, and Frase presents his arguments in terms of the traditional figures. For consistency they have been translated into Johnson-Laird's nomenclature, which is used throughout the rest of this thesis, but these are followed by the traditional figures in

highest certainty, while figure 1 (IV) was slowest. The trained group were relatively less confident on figure 1 than the untrained group.

In the second experiment groups of subjects were forced to respond within a time limit – either 20 or 30 seconds or unlimited. In this case, the items were presented on cards and subjects wrote their answers in an answer book. The results showed that the 20 second time limit group were substantially less accurate than either the 30 second or unlimited groups. Similarly, the 20 second group were less confident that the 30 second group, though the unlimited group were less confident still. There was no interaction between figure and time limits. Finally, once again figure 1 (IV) produced most errors and figure 4 (I) produced least, a result incidentally replicated in Erickson (1974) (see below).

In both these experiments

> It was also noted that Ss' reports on how sure they were of their judgements on the different figures were differentially affected by instruction, while error and time scores were not. In terms of matching their sureness to the relative distribution of errors, the trained group was more successful.
>
> (Frase, 1968. P411)

This observation is particularly interesting in the light of Evan's suggestion of a dual process account that separates the justification of the reasoning from the actual mechanisms that perform it.

Frase makes another attempt to apply associationist theory to syllogisms in **Pezzoli and Frase (1968).** This presents a review of a very complex experiment in which 5 syllogisms, invalid in both figures 3 and 4, were presented under a variety of conditions. They found no effect of varying the times between premise presentations from 2 to 10 seconds. When they varied the degree of association between the terms of the premises (ensuring that there was none between the terms of the conclusion) they found that highly associative material induced errors (i.e. the acceptance of invalid conclusions) in figure 4 (B - A, B - C), but reduced them in figure 3 (A - B, C - B). Finally, when they preceded the traditional premises by varying amounts of what they aptly term "verbiage", they found that adding nonsense context actually slightly enhanced performance (i.e. tendency to recognise invalidity).

5.4.2. Diagrams and Circles

Another experiment to allow for the effect of the figure is reported in **Erickson (1974).** In this experiment subjects were asked to complete syllogisms by supplying either a conclusion or a premise, although only the supplying of conclusions is discussed. The 24 (traditionally) valid syllogisms were presented in terms of As, Bs and Cs, with the conclusion always quantifying the relationship between As and Cs. Subjects were not given the option of (wrongly) indicating that there was no valid conclusion.

Erickson also varied the instructions that the subjects were given. In addition to a control group given minimal instructions to do the task, a second group were shown how to use what Erickson calls Venn Diagrams (although elsewhere in the paper he uses this term to describe Euler Circles) and encouraged to draw them during the experiment, while a third was given a list of rules

---

Roman numerals.

(e.g. "there is never a conclusion if both premises are negative") they could refer to. However, these instructions had minimal effect, although Erickson notes that more thorough instruction does improve performance. This suggests that the training in this instance was simply insufficient, and is best simply ignored.

Erickson suggests that syllogism solving is a three-stage process, which makes him one of the first to present his theory explicitly in terms of a process for getting a solution. In this, and indeed in the breakdown into the three stages, he is possibly following Wason and Johnson-Laird (1972 – see Chapter 6 below), although his overall approach is somewhat different. In the first stage each premise is interpreted to build some kind of set representation, which Erickson denotes by means of Euler circles (calling them Venn diagrams). In the second stage the representations for the premises are combined, while in the third the subject selects a sentence form to describe the combined representation. However, this is only a skeleton for a theory which Erickson goes on to flesh out, by specifying the details of each of the stages, in three different ways.

Stage one is the same in all of Erickson's models. Subjects do not attempt to deal with the uncertainty in the set relationship specified by the premises but simply make a (weighted) random choice among the possible relationships. Stage three, the generation of a conclusion from the combined set relationship, is also common, subjects selecting their answer by making a (weighted) random choice from those that are appropriate. If universal conclusions are preferred to particulars and the order of the terms in the conclusion is constrained (which it is when using only traditional syllogisms) there is usually only one possibility. Where there is more than one, the probabilities of each choice are influenced by the context (effectively the atmosphere) established by the premises.

What distinguishes Erickson's models is stage two, where the representations of the individual premises are combined. Many pairs of set relations can be combined in more than one way. Erickson's first, "Random Combination", model assumes that subjects make a random choice from the possible methods of combination and use that to produce a combined representation from which a conclusion is read. In contrast, his "Complete Combination" model assumes that the premise representations are combined in all possible ways, and subjects choose only conclusions that are compatible with all resulting relationships.

Erickson presents a table showing, for each premise combination that allows a (traditionally) valid conclusion, the relative frequencies of responses given by his subjects and the probabilities of those responses as predicted by each model. He stresses that the probabilities used at each stage have not been optimally chosen, and thus that quantitative agreement with results should not be expected. Nonetheless, he is able to claim that

> First, in every case but one, both models predict the most popular response, although in the case of [the EI] syllogisms, the quantitative fits are only fair. Second, both models, but especially the Random Combination Model, predict all popular responses
>
> (Erickson, 1974. P318)

Erickson then goes on to apply his models to the data of Ceraso and Provitera (1971), which include subjects both correctly and mistakenly answering that there is no valid conclusion. This leads him to point out serious deficiencies with both his theories. The Random Combination Model, since it considers only one possible combination of premises, will always be able to give a

conclusion, and thus will never predict a "can't say" response. In contrast, the Complete Combination Model will always be aware of the full range of possible combinations, and thus will never give a conclusion where none is appropriate.

Undaunted, Erickson lays his course between Scylla and Charybdis by putting forward his third model. In this case, the result of the combination of premises in stage two is determined by (suitably weighted) random choice, but with the addition of a possibility leading to the response that there is no valid conclusion. Then, on the basis of the Ceraso and Provitera data, he estimates the 18 probabilities relevant to pairs of premises that do not support a valid conclusion, and goes on to demonstrate that these values fit well with the invalid syllogism data of Chapman and Chapman (1959). Finally, he points out that the third model will fit the results for syllogisms with valid conclusions at least as well as the better of the other two models, presumably by extending the table of relation combination probabilities to include all possible premise pairs.

Dickstein (1978b) criticises Erickson's results, pointing out that while his theory gives a very good fit with the complete set of responses, if those items influenced by Illicit conversion are omitted, the correlations with actual results are greatly reduced. He then concludes that

> The model proposed by Erickson (1974) does not appear to account very well for subject performance on those invalid syllogisms that cannot be accounted for by conversion. Since these syllogisms constitute half of the total set of syllogisms, this constitutes a serious deficiency of the model
>
> (Dickstein, 1978b. P539)

In response to this, Erickson might want to say that the probabilities he offered were derived for the whole group, and their accuracy thus should be assessed on it. However, Dickstein's observation is nonetheless important, showing that illicit conversion delimits a set of items for which Erickson's theory is not accurate, though the theory itself has no means of explaining why this should be so.

The most fundamental problem with Erickson's approach, however, is that it is not the right **type** of theory – it is simply unacceptable within the current paradigm of psychology. Erickson, working within an essentially behaviourist paradigm, produced a theory couched in terms of probabilities of different decisions being taken. However, from the current cognitivist viewpoint this is simply not an acceptable level of detail. Dickstein points out this problem too, saying that Erickson's theory "appears inadequate"...

> because the various probabilities are empirically derived... and no rational or psychological explanation for them is provided. No attempt is made to explain why the probability of selection of a particular combination will exceed the probability of selection of a different combination, or why the probability of recognition that there are incompatible possibilities is greater for one set of premise interpretations than for another
>
> (Dickstein, 1978b. P 539)

These were simply not the kind of problems that Erickson's (paradigm of) psychology suggested he should investigate.

5.4.3. Backward Processing

In addition to criticising Erickson, Dickstein also presents his own theorising. In **Dickstein (1978a)**, possibly influenced by Wason and Johnson-Laird (1972), he presents an extension of Chapman and Chapman's Illicit Conversion that tries to take the figure into account. The theory is

supported by two sets of experimental results: the responses from the "standard instruction" group from Dickstein (1975) (see Section 5.3.2) and the results of a new experiment in which different groups of subjects tackled each syllogistic figure. In each case symbolic material was used and subjects selected their response from a list that included the four possible conclusions and a "no conclusion" option. Dickstein presents the complete table of the response distributions from each experiment, which, depressingly, is the first time in nearly forty years of research that anyone had produced and published a complete set of responses.

Dickstein begins by pointing out that, while traditional syllogisms only allow one ordering of the terms of the conclusion, it is often possible to draw a different conclusion if the terms are in the opposite order. He suggests that this *backwards processing* is one of the factors affecting performance in dealing with syllogisms, leading subjects to accept invalid conclusions because they are valid "in the other direction". In essence, this is the application of Illicit Conversion not only to the premises, but also to the conclusion, and is reminiscent of the "different" category proposed by Revlis (1975) (see section 5.3.1). Dickstein goes on to suggest that the tendency towards backwards processing will be influenced by the figure of the premises presented. In particular, in (Johnson-Laird's) fourth figure both the premises run "forwards", so there should be minimal tendency towards backwards processing, while in the first figure, both the premises run "backwards", so the effect should be strongest.

These suggestions lead him to make several predictions. On those syllogisms that support different conclusions in "forward" and "backward" modes, accuracy should be best in figure 4 and worst in figure 1, and this should be accompanied by a corresponding increase in the frequency of the "backward processed" response. On the other hand, where the same response results from both modes, the figure of the problem should have no effect.

Dickstein presents statistics based on his results which illustrate the effect of figure that backward processing predicted. In doing so, he stresses that it is vital to discount items that are affected by Illicit Conversion, the effect of which he demonstrated in Dickstein (1975). He goes on to argue against the associationist theory proposed by Frase (1968), which predicts that the arrangement of terms should influence all the syllogisms in each figure equally. Dickstein, however, finds that there is a significant effect only where forward and backward processing give different results. He also comments that while the data of Erickson (1974) appear to support this theory, this is no longer the case when the results of problems influenced by Illicit Conversion are discounted.

Dickstein (1978b) begins with the attack on Erickson's behaviourist theories mentioned above, and then goes on to present his own account in "high level" terms that are reminiscent of Henle. Dickstein (1975) is deemed to have established the case for Illicit Conversion, so only those invalid syllogisms it cannot account for need be considered. These 32 problems are broken into three groups, for which three separate error mechanisms are postulated. One group comprises pairs of negative premises, where

> It is proposed that subjects err by failing to consider the possibility that the subject and predicate may be related without the mediation of the middle term
>
> (Dickstein, 1978b. P540)

The second group has pairs of particular premises, and

> subjects err here by assuming that the M referred to in the major premise and the M referred to in the minor premise are the same
>
> (ibid)

This is a mechanism Henle(1962) explicitly proposes, and which predicts the same conclusions as the Atmosphere Effect. The third class contains only **IE** premises, which have no traditionally acceptable conclusion. Errors here are explained by either the mechanism of the "double negative" group, or by the mechanism of "backward conversion" argued for in (Dickstein 1978a).

Dickstein makes two predictions on the strength of this diversity of mechanism: As usual he argues that the different mechanisms will mean that the groups will differ in difficulty, but he goes on to suggest that errors produced by the same mechanism will be correlated – predictions which neither Atmosphere Effect nor Erickson can explain. He analyses the results from the second group of subjects from the experiments of (1978a), and finds that the different groups have clearly distinct difficulty (accuracy), that the predicted correlations are high and the others are insignificant. The results for the other group are similar, but by no means so clear cut.

Finally, Dickstein states that he believes Johnson-Laird's experimental paradigm is superior to his own. Presumably, therefore, he must have been aware of the strength of the figural effect that they reveal, but he makes no attempt to account for them.

5.4.4. Newell and Search Spaces

The discussion of the desirable features of a theory of syllogistic reasoning in Section 4.4 mentioned that **Newell (1981)** severely criticises the tendency to study cognition in a haphazard way, without "essential orderliness" or a clear task-independent theory. In the light of such views, it is no surprise that Newell has his own candidate for the job, and since syllogistic reasoning is one of the tasks that he uses to illustrate its application, it is appropriate to consider his ideas here.

Newell's proposed task-independent theory is based on the idea of the *problem space*, which is argued for and described in more detail in (Newell and Simon, 1972). A problem-space is a means of representing states of the world and a range of operators for creating or reaching one state from another. A particular problem is the specification of an initial and a goal state, a solution is a sequence of operations for achieving one from the other and finding it is a matter of searching the problem space. This is a cycle of repeatedly choosing a state and an operator, applying it and evaluating the resulting state. The application of the operators can be arbitrarily complicated – Newell offers integrating an equation as an example – and may involve the exhaustive search of a different problem space. In contrast, the control of the search process – the first and third steps – can only rely on immediately accessible knowledge, and the decisions taken will determine the way the problem is solved.

Everybody is assumed to have some basic rules for guiding the search of a problem space: e.g. "don't make a move that undoes the last one you made", "never move to a state you remember being in before" (memory for previous states is, obviously, limited) or "consider the last new

position you created", which results in a depth-first search. Newell also suggests that they might well be able to adopt some A.I. search control strategies such as Hill Climbing, Means-Ends Analysis and Operator Subgoaling, which he terms *weak* methods because they are domain independent. However, as subjects becomes familiar with the task, they will develop more powerful search control strategies that are specific to the problem space. Skilled behaviour is what happens when the goal selection process is sufficiently accurate that search is avoided.

To illustrate this approach, Newell presents accounts of problem solving (the "Towers of Hanoi") and syllogistic reasoning. He suggests that the problem space for syllogistic reasoning is based on descriptions of representative things that might exist, to which he tries to lend credence by presenting a "monologue"[7] of syllogistic reasoning in these terms. The representative entities are tagged as being either with or without the properties involved in the syllogism, and as representing something that necessarily or possibly exists (or, by their omission, definitely non-existent). These are manipulated by three types of operator.

(1) Operators for interpreting the premises by adding the relevant possible and necessary entities to the problem state. E.g.

All A are B ==> necessary(A+ B+), possible(A- B-), possible(A- B+)

Notice that there is no "possible(A+ B-)", which means that it cannot exist.

(2) Operators for combining objects, provided that the corresponding attributes agree. E.g.

possible(X Y), possible(X Z) ==> possible(X Y Z)
necessary(-A Y), possible(-A Z) ==> possible(-A Y Z)

There are also operators for inferring that possible objects are necessary from the absence (impossibility) of "complementary" objects.

(3) Operators for giving possible conclusions when they recognise necessary entities with properties relating both the end terms.

The solution of a syllogism proceeds by the sequential application of operators to problems states which record all the currently known restrictions on the set relations described by the syllogism.

This theory obviously has many similarities with Johnson-Laird's account based on mental models, which forms the focus of the next chapter. Both involve a few entities tagged with properties and each representing the whole class of things which have that combination of properties. These entities may definitely exist, or be only possible (cf. bracketed items), and can be marked as explicitly not having a property (barriers or negative links).

However, Johnson-Laird (1983) acknowledges none of this similarity. Indeed, he minimises it by presenting Newells theory (P89 - 91) as a "symbolic variation" of Venn diagrams.[8] While doing so, he criticises it for representing impossibility by the absence of entities, which he believes makes it indistinguishable from ignorance. However, this is unwarranted, since uncertainty is

---

[7] Probably not an actual protocol, but invented. Note that offering a (mock) protocol, and indeed the fact that he is mentions syllogistic reasoning at all, suggest that unlike many others, including Johnson-Laird, he sees the task as an explicit exercise in deliberate problem solving.

[8] As opposed to Euler circles. Each of Newell's entities corresponds to one area of a Venn diagram. A missing entity corresponds to the region shaded as empty, while "necessary" and "possible" entities respectively correspond to regions marked as non-empty and not marked at all.

revealed by considering the problem and arriving at an entity which is present but not "necessary". These "possible" entities represent uncertainty in just the same way as Johnson-Laird's own bracketed entities. He also criticises it because "As Newell correctly points out... the latter convention [omitting impossible entities] makes the notational system vulnerable to errors of omission"[9] (Johnson-Laird, 1983. P 90). It is not clear what either of them mean by this. Given his view of the simplicity of the syllogistic task (see below), Newell might well be concerned that a subject's failure to do something (i.e. include an entity) can result in a sensible problem state, and thus lead to illogical conclusions. But Johnson-Laird admits the importance of explaining reasoning errors, and even faults Newell for not doing so. Moreover, these omissions seem indistinguishable from Johnson-Laird's ascribing of subject's errors to the omission of bracket items, or the failure to pursue model testing to completion.

Johnson-Laird's main, and justified, criticism of Newell's theory is that it fails to account for subjects' reasoning errors. This leads Johnson-Laird (1983, P91) to describe the account as "an illustration of an approach... rather than a fully-fledged theory", which is precisely what Newell says it is, and to dismiss it as "plainly both descriptively and explanatorily inadequate" while "there is an alternative and simpler theory on the same general lines" (Johnson-Laird, 1983. P91). This is a reference to the account he then presents based on using the premises to control the deletion of lines from a table of all possible entities, which he himself then rejects as implausible. However, in mentioning that this stand-alone theory is simpler than Newell's problem-space account, Johnson-Laird seems to be completely missing or ignoring Newell's point. It is **crucial** that any task-specific theory can be seen as part of some grander scheme. For Newell, this is the idea of searching problem spaces, and the whole purpose of presenting the account of syllogism solving was to illustrate the kind of way it might be applied (by some subjects, sometimes) to a reasoning task.

Nonetheless, the inability to deal with errors is, as Johnson-Laird points out, a fatal flaw in the account. Newell nods in this direction by suggesting that the illicit conversion of an **A** premise would result from an error in a "secondary" object, rather than a "focal" one. However, he never clarifies what this distinction amounts to, and in any case most interpretations suggest the conversion of an **O** depends on a "focal" object! However, Newell tackles syllogistic reasoning solely in order to illustrate the kinds of theories that searching a problem space can suggest and in doing so simply does not consider errors an important feature of syllogism performance. The reason why this is so becomes clear when, having presented the two example task domains (syllogisms and "Towers of Hanoi"), Newell goes on to discuss the way they illustrate the implications of the idea of searching a problem space.

---

[9] Newell, 1981, page 708. In fact, Newell suggests that it is the **problem space** that is vulnerable, but in the context of different subjects using different problem spaces at different times.

> Two conditions at least permit the structure of behaviour for a task to be obtained from just the general problem space structure. In one, the task is simple and transparent, relative to the cognitive ability of the subject. Then elementary search control knowledge suffices. This is the situation in syllogistic reasoning tasks. In the other, the situation is novel, so that the subject does not have much prior knowledge and does not have time to extract much new specific knowledge. Then the subject must rely on elementary knowledge. This condition produces novice problem-solving behaviour.
>
> (Newell, 1981. P713)

Since Newell sees the syllogistic reasoning task as "simple and transparent, relative to the cognitive ability of the subject", he has no reason to suppose that the errors it induces will be particularly interesting, and thus he pays them little attention. Since subjects overall error rates are round about 40%, rising to over 90% for some of the harder syllogisms, this is clearly a use of the phrase "simple and transparent" with which many people may not be altogether familiar.[10]

Allowing for Newell's unusual assessment of the difficulty of syllogistic reasoning, he is suggesting that the mechanisms for searching problem spaces will reveal themselves when the situation is either very simple or completely unfamiliar. Recalling that Newell believes that "skilled behaviour is when operator selection is sufficiently accurate that search is avoided", expertise must be seen as additional operators. These could be of two forms, either making new movements through the problem space, or recognising and making explicit the applicability of other operators. While this expertise is being acquired, the subject might well employ incorrect notions about the problem space. This could lead to arbitrary, and thus inexplicable, and inappropriate behaviour, which is why intermediate problems may not reveal the behaviour associated with space searching. However, when the task is new, the subject will have no wrong ideas about how to tackle it and will once again exhibit behaviour suggested by "elementary search control knowledge".

What do Newell's ideas say about the processes of syllogistic reasoning? The illustrative account he presents is of little value per se, since it makes no mention of reasoning errors, which abound in every experiment. This is because Newell categorised syllogistic reasoning as a simple task, and thus, like so many others before him, concentrated on explaining correct reasoning in his preferred way. However, simple search control knowledge can only do this on the very strong assumption that subjects will employ styles of representation and manipulation that reflect a sound grasp of the problem. Specifically, they must recognise that syllogistic problems only ever involve determining the status of eight types of entities (or seven if one ignores the background provided by the vast majority of the universe, which has none of the properties involved). However, it is questionable whether most subjects actually recognise this – not least on the basis of their syllogistic abilities. Such recognition could only come as a result of meta-level reasoning: reasoning not about a syllogistic problem, but about syllogistic problems in general. However, as pointed out above, this is not the way people tend to tackle their problems or, thankfully for psychologists, psychology experiments. Thus there is little reason to favour such a style of

[10] It is possible to suggest that Newell is thinking of the reasoning behaviour of (logically) trained subjects. If this were the case, the the mismatch between his ideas and the performance of untrained subjects would be understandable. However, he says nothing to suggest that this is so or to argue for the possibility of generalising from such a specialised situation.

representation, or the view of syllogism solving that it suggests.

However, Newell stresses that this approach is only suggestive, and that he is mainly concerned with the application of the idea of space searching. Therefore, some attempt must be made to apply the approach to syllogistic reasoning, even if it is recognised as far from "simple and transparent". Unfortunately, this is difficult, because experimental subjects are not skilled at the task, at least in terms of getting valid responses. Thus it is at neither of the extremes – triviality or accomplished skill – where space searching will be apparent. It is firmly in the "muddle in the middle", where the effects Newell hopes for might well be masked by subject's ill-conceived ideas of what they are doing. This is certainly the picture suggested by the modelling of individual subjects, to be described in Chapter 7, where although their overall grasp of the problem was sound, their knowledge of how to go about solving it was lacking. As a result, even though the idea of the problem space might be a valid and powerful inter-domain theory, and thus sorely needed, its power to restrict theories of syllogistic reasoning is minimal.

Although the theory of syllogistic reasoning that Newell presents has many obvious and serious shortcomings, it is nonetheless important. Unlike most other theorists, he makes clear how his account fits in with and bears upon the explanation of other cognitive processes. Moreover, whereas all other theorists have assumed that the mechanisms that underpin performance on the task are those that support language understanding or everyday reasoning, Newell has proposed that it has much closer ties with formal problem solving.

5.5. Summary

The extensive psychological research into syllogistic reasoning presents a depressingly chequered and blurred picture. Experimenters have continually varied almost every conceivable feature of the paradigm, and have produced results for almost every conceivable form of syllogism-related experiment – validation versus selection, symbolic versus familiar material, valid versus invalid forms. Unfortunately, this variation has reached the point where no two have carried out identical, or even closely related, experiments, and there are few places where related experiments can be directly compared. Nonetheless, they have served to highlight the importance of a range of features. Evidence from varying instruction forms shows that subjects errors arise from more than one source. Similarly, the use of simpler premises suggests that one of these is related to a tendency to illicitly convert the premises offered, as does the precision with which illicit conversion predicts subjects responses for the syllogism forms where it makes a difference.

Of the theories offered to explain these varied results, only Newell's constitutes the kind of information processing account that cognitive psychology seeks. Moreover, in terms of Johnson-Laird's proposed criteria for theory evaluation, none deal with individual differences, the acquisition of the relevant abilities or the development of formal logic. Most achieve a reasonable coverage of data, in the sense of predicting the most common errors or certain gross trends in difficulty, though Erickson has done so mainly by pushing the observed response distributions down into internal probabilities. The potential for correct reasoning has, in the form of a desire for a "rational" mechanism, proved something of a tar-baby. Reasoning errors have essentially been accounted for in terms of the adoption of unsound rules of inference, and there has been much

debate concerning the relative "rationality" of various strategies. However, the vital matter of the motivation of their adoption – the distinction between deliberately trying to bend the rules and knowing no better – is not tackled.

Finally, the theories offered have, for the most part, been uniformly rootless, and justifiable targets for Newell's remarks concerning the writing down of flow charts. The exceptions are Frase's associationism and Newell's own "problem space" account. The former seems seriously deficient in the picture it paints of the subjects' grasp of the task and the meanings of the quantifiers involved (i.e. they are not mentioned). In contrast, the latter grants them a full understanding of the situation, but seriously underestimates the difficulty untrained subjects have in reasoning soundly. As a result, it can only be viewed as an example of how Newell's approach can be applied, which is, of course, what it was meant to be.

Clearly, these accounts leave something to be desired. The next chapter focuses on the work of Johnson-Laird, who offers a processing account with considerably greater appeal than those discussed so far.

CHAPTER 6

Johnson-Laird's Account of Syllogisms

6.1. Johnson-Laird's Experiments

Johnson-Laird first addressed the task of syllogistic reasoning in **Wason and Johnson-Laird (1972)**. As well as forming the basis of his own subsequent theorising using mental models, the ideas presented there appear to have exerted a detectable influence on both Erickson and Dickstein (see Chapter 5). They are formulated on the basis of an experiment in which subjects were asked neither to validate a given conclusion nor select one from a list, but to give their own. The experiment was also a break from tradition in that the items presented were the 27 premise pairs that would support a conclusion in at least one direction, so every item had a valid conclusion, and the subjects knew this. Finally, the syllogisms were presented not with symbolic material but with "everyday" words.

In the light of the fate that befell Henle's attempt to introduce meaningful material into a reasoning experiment, the decision to use everyday words may seem to need justification. For Wason and Johnson-Laird, as with so many before them,

> The main point of the experment, of course, was to see to what extent the subjects' error were compatible, or incompatible, with the Atmosphere effect
>
> (ibid, P137)

They were well aware of how strongly the material in the syllogism had influenced subjects in Wilkins (1928). However, they point out another feature of those same results: namely that for several problems where the Atmosphere Effect is strongly supported by results using symbolic material, "familiar" material gives completely different results. Thus Wason and Johnson-Laird chose to launch their attack on the Atmosphere Effect where it had already been seen to be weak. Like Wilkins, but unlike Henle, they attempted to introduce familiar material without introducing any content. They exemplify this familiar nonsense with "all the nuns are nurses, none of the nurses are midwives", which they describe as "neither contentious, nor likely to arouse questions of truth or falsity" (P137).

Wason and Johnson-Laird do not present the complete set of their results, but instead highlight features to make specific points. The first thing they try to do is to show that their data do not support the Atmosphere Effect. Classifying problems by the mood of the two premises and that of the strongest valid conclusion (but ignoring the figure), they produce a table to show the numbers of error responses that were compatible and incompatible with the Atmosphere Effect. They conclude that

> There were reliably more errors which were incompatible with the atmosphere ... than were compatible. But what the results show is that the tendency to conform to atmosphere varies a great deal from one sort of problem to another.
>
> (Wason and Johnson-Laird, 1972. P137)

What they do not point out is that in the problems that give very few errors compatible with the Atmosphere Effect, the correct conclusion is compatible with it. It was pointed out in Chapter 4 that Wason and Johnson-Laird are unlike most other theorists in treating the Atmosphere Effect the way Woodworth and Sells intended: not as a complete theory of syllogistic reasoning, but as a mechanism that operates when the problem is beyond the subject's abilities. Here, however, they are making the complementary error of assuming that, if the subjects give the right conclusion, the Atmosphere Effect was not involved. As Evans (1982, P86) points out, the fact that subjects produce a sound conclusion does not show that they have done so by sound methods.

Having criticised the Atmosphere Effect, Wason and Johnson-Laird introduce their own theorising by suggesting that syllogism solving can be seen as a two-part process: first the premises are interpreted, and then they are combined. They then proceed to examine each of these parts in turn.

While discussing the interpretation of the premises, they describe the experiment conducted by Jean Waddington and outlined in section 4.2.2 above. This showed that even when subjects were instructed to ignore the content words, the range of situations they consider a sentence describes is none the less strongly influenced by them. Then, citing the inspiration of Inhelder and Piaget's thoughts on the treatment of quantifiers by young children, they go on to suggest a mechanism by which an adult interprets (universal) syllogistic premises:

> He tends to interpret the statement, "all X are Y", by setting up a representation of the class of instances corresponding to X, and tagging each instance with the attribute Y.
>
> (Wason and Johnson-Laird, 1972. P148)

This is clearly a description of what Johnson-Laird will subsequently call a mental model. They then go on to suggest that with abstract or symbolic material, this model is taken as the whole universe under consideration. This is the same error as they point out Inhelder and Piaget (1958) observe in children, and creates a situation in which conversion of an A premise is valid. However,

> In the case of meaningful or familiar material, however, knowledge of the world will prevent the individual focusing solely upon the quantified term.
>
> (ibid, P149)

In other words, the situation the subject is considering will not support the conversion of a premise, and thus the observed rate of errors from Illicit Conversion is reduced.

Turning to the actual combination of the premises, Wason and Johnson-Laird argue that

> First, syllogisms differ widely in their difficulty. There are some which are straightforward and which most intelligent adults can solve in a few seconds. There are others, however, which are extremely difficult. This is soon appreciated by the reasoner, yet even after several minutes of thought he may still produce a fallacious answer. Second, even the easiest of syllogisms tends to take a considerable time to solve in comparison to a three term series problem.
>
> These two facts imply that the process of combination is essentially a series of processes rather than a single act
>
> (ibid, P151)

Possibly the fact that there is a continuous scale of difficulty is to be taken as evidence against a

mechanism that either solves the problem or does not. Otherwise it is not clear why this fact is relevant to whether they are solved by a series of processes: Some rivers are significantly easier to jump across than others, yet this does not obviously imply that river-jumping involves a series of processes.

Wason and Johnson-Laird also make a very telling observation on how someone often goes about solving a syllogistic problem:

> On presenting the premises, say:
>
> > None of the musicians are inventors.
> > All the inventors are professors.
>
> an articulate individual is likely to announce an obviously tentative answer after a relatively short time:
>
> > None of the musicians are professors.
>
> If one then asks "are you sure?", or merely refrains from comment, a considerable amount of time will probably elapse during which it is clear that intense cogitation is going on. A series of successive approximations, culminating in the correct conclusion, will probably be forthcoming, e.g.:
>
> > None of the professors are musicians.
> > None of the professors who are inventors are musicians.
> > Some of the professors are not musicians.
>
> ... Not until the final answer is the process completely carried out.
>
> (ibid, P152)

The idea that conclusions are generated by successive approximation thus becomes central to Johnson-Laird's subsequent theorising. Here, with Wason, he says:

> It seems that the reasoner initially constructs his first approximation to the answer, working in an intuitive fashion, and then submits is to a series of logical checks. Such checks probably involve trying out various ways of combining the information in the premises. They will sometimes lead to a revision in the answer, and even perhaps to the decision that no definite conclusion can be derived from the premises.
>
> (ibid)

On the subject of how the initial conclusions are generated, Wason and Johnson-Laird suggest:

> It is plausible that the individual, untrained in formal logic, appreciates that if there is a "restricted" premise, concerning only **some** members of a class, then the conclusion must be similarly restricted. Likewise, he appreciates that if there is a negative premise, the conclusion must also be negative. (This second principle does admit exceptions, however, since "some A are not B" is often taken to imply "some A are B".) These two principles, of course, correspond exactly with the rules for the atmosphere effect. But in this case they are reflections of an underlying logical knowledge of what properties of premises are transmitted to conclusions, rather than recipes for the purely superficial matching of verbal elements.
>
> (ibid, P153)

The (lack of) content of this last disclaimer was suggested when the ideas of Henle were discussed in the previous chapter. Contrary to what Wason and Johnson-Laird would wish, this quotation simply highlights a sense in which the Atmosphere Effect can be seen as part of a highly rational procedure, provided it is used in combination with some other "logical" mechanism, which, of course, Woodworth and Sells explicitly admit that it is.

Unfortunately, Wason and Johnson-Laird's theory is deficient, in that it cannot account for one of the most striking features of their results, Since subjects are, for the first time, generating conclusions, they are free to order the terms within them in either of the two ways. Whenever a conclusion of the form "some (or no) X are Y" is valid, so also is a conclusion of the form "some (or no) Y are X": they are logically identical. However, Wason and Johnson-Laird found that their subjects had a strong preference between the two forms depending on the figure of the premises. With premises in the first figure (i.e. A - B, B - C) subjects prefer to offer conclusions in the form A - C, while in the fourth figure (B - A, C - B) they prefer C - A. In the remaining "symmetrical" figures, subjects show no preference. Subsequent work by Johnson-Laird and Steedman suggests that this phenomenon, which they term the *Figural Effect*, is reliable, and equally observable in both valid and invalid conclusions offered.

In order to get around this omission, Wason and Johnson-Laird consider an extension of the theory proposed by Hunter (1957) to deal with a similar kind of bias observed in three-term series problems. Carrying his theory across to syllogisms, they suggested that subjects determined the term order in their conclusions by arranging the premises into a "natural" order, so that the middle terms were together,[1] at which point they were simply "expunged". Thus premises in the first figure (A - B, B - C) would immediately give a conclusion A - C, whereas those in other figures would need to be manipulated, either by reversing the order of the premises or by conversion, possibly illicitly.

Having proposed this mechanism, they proceed to undermine it by pointing out that their data do not support its predictions. According to such a theory, a **4IA** syllogism would be solved by (validly) converting the first (I) premise, to yield a **1IA** syllogism which would then immediately yield an **I** conclusion linking A - C. This would suggest that the subject's responses for both the **4IA** and **1IA** syllogism should be dominated by responses of this kind. However, while this prediction is born out for the **1IA**, the **4IA** is dominated by **I** conclusions linking C - A, which argues against a theory suggesting that the **4IA** is solved by converting it to a **1IA**. Similar evidence against the any theory based on converting the second premise comes from the **1AE** and **3AE** syllogisms.

As a result, Wason and Johnson-Laird abandon mechanisms based on Hunter's work. Instead, they try to ascribe the effect "in purely descriptive terms" to the fact that the conclusions are generated in line with a (disconcertingly complex) rule. This states that the term to be quantified over in the conclusion is the term quantified over in the restrictive or negative premise, if it is an end term, otherwise it is the quantified term in the other premise. However, it is not clear what "purely descriptive" amounts to, and whether subject's obedience to this rule is as a result of explicitly following it, or emerges from the operation of a combination of mechanisms.

Turning finally to the actual logical checking of potential conclusions, Wason and Johnson-Laird offer only a very rough outline of a mechanism. They use Euler Circles to represent the situations the way the subject is considering them (cf Erickson) and suggest that various

[1] Johnson-Laird (1983, P70) points out that this is the form that Aristotle considered "perfect".

combinations of the premises are considered not with the aim of falsifying a potential conclusion, but of verifying it.

Johnson-Laird's next discussion of experiments on syllogistic reasoning is in **Johnson-Laird and Steedman (1978)**, which includes the description of three pieces of work. The first is the experiment described in Wason and Johnson-Laird, although it is worth noting that there are slight differences in the presented results. For every result where Wason and Johnson-Laird report a number of subjects giving a certain response, Johnson-Laird and Steedman have one or two subjects more (though never less) in the corresponding category. The most likely cause of this is a (possibly unintentional) reclassification of some answers which were not presented in purely syllogistic form (see 6.2.1.1 below).

The second experiment is in some sense the obvious "completion" of the first, with subjects being presented with all 64 possible premise pairs in a random order, and asked to write down their own conclusion if they thought there was one. The subjects were 20 students at Columbia University,[2] and once again, familiar but meaningless materials were employed and subject's thinking times were measured, probably using a stop-watch. Each subject performed the experiment twice, with an interval of one week between them. The group results from both "sittings" are presented, in the form of a table of distributions of responses to all 64 problems, and henceforth they will be referred to as Steedman1 and Steedman2. However, attention is concentrated on the results obtained when subjects took the test for the second time, on the grounds that the patterns of results are "very similar". However, no justification is offered for studying the second sitting in preference to first, and while the authors note that 19 out of 20 subjects were more accurate at the second sitting, even though they (presumably) received no feedback on their performance, they say nothing to explain this.

Turning to the results of these experiments, Johnson-Laird and Steedman use statistics related to subjects' correct responses to demonstrate that both figure and premise mood are significant and strongly interacting factors. They point out that the first figure significantly promotes conclusion drawing, and report the discovery, in both experiments, of what they term the *Figural Effect*, which was first observed in the results of Wason and Johnson-Laird, but is now shown to appear whether or not the conclusion is valid. It can also be seen as a statistically significant bias in the total numbers of conclusions drawn in each direction in a given figure – that is, it even affects those conclusions that are not symmetrical.

Johnson-Laird and Bara (1984) present three experiments conducted by Bara, in each case using 20 students from Milan University as subjects. The first experiment is very similar to the main experiment reported by Johnson-Laird and Steedman, but following up an idea first suggested in Wason and Johnson-Laird (1972, P155). Instead of being allowed to think for as long as they liked, subjects were obliged to respond within 10 seconds, and these time-limited group responses are tabulated in Johnson-Laird and Bara. When subjects had answered all 64 syllogisms, they were

[2] The experiment was actually supervised by Janellen Huttenlocher. However, this is her only direct contribution to the work being discussed, so for the sake of simplicity, therefore, the experiments will be referred to as Steedman's.

given a five minute break and then re-presented with the same problem items, together with their initial conclusion, and given a full minute in which to reconsider their response. Knowing that they would be allowed to change their minds later was intended to make subjects less inhibited about offering conclusions before they were ready (i.e. sure). No indication is given of the manner or degree of subjects revisions.

The general pattern of results of this experiment were very similar to those reported by Johnson-Laird and Steedman, with one or two conclusions being clearly preferred for most problem forms, and the Figural Effect is universally and significantly present. However the imposition of a ten second time limit had the dramatic (and for Johnson-Laird surprising) effect of increasing subjects' tendency to suggest that there was no valid conclusion. Moreover, the increase in this tendency was affected by the figure of the syllogism.

The second experiment that Johnson-Laird and Bara report tries to assess the generality of the mechanisms responsible for the Figural Effect by attempting to detect it in a non-syllogistic problem. Specifically, they use a three term series problems denoting kinship between individuals (e.g. "Fred is related to Bert, Bert is related to Mary, therefore Fred is related to Mary"). The position of the terms within the premises allows the syllogistic figures to be identified, though the problem domain is simpler because there are only two possible premise forms (i.e. "is related" and "is not related"), thus giving only four possible problems in each figure (as opposed to syllogistic problems 16 per figure). Ten undergraduates at Sussex University were given examples of all 16 possible problems, and invited to draw their own conclusions. Johnson-Laird and Bara found a general preference to use the terms in the order they are presented in the premises, but set against this all the features that constitute the Figural Effect were clearly present.

The final experiment that they present is almost an exact replication of the main experiment of Johnson-Laird and Steedman – i.e. all 64 possible problems, free response times and (Italian) "familiar nonsense" material – and as before, a table of the group responses (i.e. number of subjects giving each response to each problem) is given. Once again the general pattern of results is similar, and the figural effect is unmistakably present. For some premise pairs subjects' preferences in the two experiments correlate well, while in others there is a wide disparity. The most widespread of these is the appearance in Bara's data (both free response and time-limited) of what Johnson-Laird terms *Gricean* conclusions which are not present at all in Steedman's. This term is used to cover a range of conclusions that were drawn in response to premise pairs that included an **O** ("some X are not Y") premise but are both invalid and inexplicable under any of the main theories of syllogistic reasoning. They can, however, be seen as much more reasonable if the **O** premise is interpreted as implying a corresponding **I** ("Some X are Y") premise. Johnson-Laird suggests that this is an example of the kind of Gricean implicature mentioned in 4.2.2: "If no X were Y, the premise would say so. Since it only says that some X are not Y, then the other Xs must actually be Ys". This will be discussed further after Johnson-Laird's theorising has been presented.

Finally, in addition to the established figural effect, Johnson-Laird and Bara observe one further feature of the results of the experiments they report:

In the case of the two symmetrical figures:

A - B	B - A
C - B	B - C

there was an interesting tendency: Where the conclusion was in the same mood as just one of the premises, the end term of the premise tended to play the same grammatical role in the conclusion as it did in the premise itself. For instance, with premises of the form "all the A are B, some of the C are B", a conclusion containing the quantifier "some" tended to take the form "some of the C are A", whereas with premises of the form "some of the A are B, all the C are B", a conclusion containing "some" tended to take the form "some of the A are C".

(Johnson-Laird and Bara, P22)

This is also observable in the results of Johnson-Laird and Steedman and Inder (see below). Note also that if applied to the asymmetrical figures (1 and 2), this "tendency" will give the same predictions as the figural effect. Indeed, the theory to be proposed in Chapter 7 accounts for both regularities as arising from the same source.

There are some methodological flaws with these experiments. Firstly, in all the syllogistic experiments subjects were told "They would be given pairs of statements about people whom they were to imagine as assembled in a room" (Johnson-Laird and Steedman, 1978. P67). Bear in mind that Johnson-Laird was intending to use the results of the experiments in order to support his idea that subjects solve syllogisms by making an internal representation of the sort of situation involved. In the light of this, telling subjects that they were "to imagine [people] as assembled in a room" could plausibly be seen as suggestive. In fact, results from replications of this experiment (see below) suggest that these instructions are not a significant factor.

Secondly, common sense, reinforced by Wilkins's (1928) results, suggest the choice of the actual material with which the syllogisms are presented is an important factor. However, this factor is neither controlled effectively nor randomised. The material in the experiment that Johnson-Laird and Steedman report consisted of 64 triplets of noun-phrases, each containing one profession or job and two pastimes, hobbies or characteristics. A minimal attempt was made at controlling for the effect of material by compiling two lists of examples, arranging the triplets so that the material used for the valid syllogisms in one set was used in invalid syllogisms in the other. Each subject did the test once with each list, "in a counter-balanced pattern". No indication is given of the extent of the differences between the two sets, with the results being simply combined, presumably on the assumption that any effect would cancel.

Bara seems to have made no attempt at controlling the influence of the material, but simply used 64 triplets, one being used to present each syllogism form to all twenty subjects. These triplets were apparently chosen with the intention to "minimise semantic relations between the terms within each premise pair while retaining plausibility for any possible conclusion, valid or invalid" (Johnson-Laird and Bara, 1984. P20). Once again the test items contained words relating to professions and hobbies, but, curiously, each triplet was arranged to use a hobby or similar as the middle term and two professions or occupations as the end terms. Since subjects are probably aware that most people have only one job, arranging that all conclusions to be considered concern the existence of people having two is not an obvious way of meeting the stated objective of "retaining plausibility".

Predicting the effect of the material employed is extremely difficult, if not impossible, as is illustrated by the following example. Subjects who had just finished a syllogism experiment (reported below) were asked whether they had been distracted by the materials in any of the examples. One subject said that she had found one particular example, relating "managers" and "guitarists", particularly misleading (although she claimed to have been able to ignore the influence anyway). Now "manager" could describe the person in charge of a pub or shop, the head of a football club, the boss of an office or the guy in charge of a workshop. Similarly Julian Bream, Johnny Cash and Jimmy Page are all guitarists. Thus one might think that the terms meet the criteria for familiar nonsense – there is no particular reason for thinking that managers would be either particularly likely or unlikely to be guitarists. However if, like the subject, one interprets "manager" as "pin-stripe suited (bank) manager" and "guitarist" as "player in a heavy metal band", then clearly one is highly sceptical that the two will go together.

This may well have a strong effect on the reasoning process, but this effect will vary according to which syllogism is presented containing the material. Lexical items that hamper the valid solution of one problem might have the opposite effect on another. If the subject is considering a positive conclusion to the syllogism, such as "some managers are guitarists", the effect would be to make them question their reasoning more carefully in an attempt to defeat it, using real-world knowledge (probably subconsciously) to guide the search for counter-examples. The subject might or might not be aware that this was happening. On the other hand, if the subject had no particular conclusion in mind, the bias in the material might well cause the premise models to be combined so as to support a negative conclusion where otherwise they might not. To give an extreme example, it seems plausible that a conclusion such as "none of the pensioners are weight-lifters" would be decidedly easier to find than "none of the artists are beekeepers". What is more, unless the subject were specifically warned in advance to look for this kind of effect they would not be likely to be aware of the influence of the material.

Finally, it should be stressed that the suggestion that subjects are being influenced by the material does not in any way force one to accept Henle's ideas about subject's refusing to accept the logical task. Subjects can both know that an always-valid conclusion is required and have adopted methods suitable to finding one (i.e. as accepting the logical task), and yet still be influenced by real-world knowledge imported into the abstract situation.

6.2. New Syllogism Experiments

In order to obtain data on the performance of individual subjects, needed to support the theory to be advanced in Chapter 7, experiments within Johnson-Laird's paradigm were conducted by the author. However, attempts were made to avoid the methodological shortcomings mentioned above, and to further explore the experimental paradigm and extend it in a number of ways.

(1) A computer terminal was used for presenting problems and accepting responses. This has three main advantages over methods employing cards and stop-watches. Firstly, the environment is completely consistent for each subject, with no possibility of the experimenter (subconsciously) influencing or hurrying subjects. In particular, the consistency of the environment can be maintained between experimenters. Secondly, there is no difficulty in

making each subject's test unique, and reliably tracking which items were used for eac subject. Finally, the results from the subject are directly available for machine processing, eliminating both the effort of transcribing them and the potential for error that this involves (this applies particularly to group B: see below).

(2) The potential importance and great difficulty of assessing (let alone controlling) the semantic influence of the content of the syllogism has already been pointed out. For this reason, the lexicalisation of the test items, together with the order of presentation, was randomised for each subject.

(3) As a result of discussions within a working group within the School of Epistemics (now the Centre for Cognitive Science) at the University of Edinburgh, it was thought likely that subjects were using the "there is no valid conclusion" option when they felt they could not really solve the problem. To attempt to discriminate such a use from a worked-out conclusion, subjects were asked to assess the confidence they had in the answers they gave.

Finally, since the new theory under investigation suggested that a subject's performance would be changed simply by doing many syllogistic problems, some subjects performed the basic experiment several times, each time solving a different set of the 64 possible syllogisms.

6.2.1. Procedure

The subjects were twenty-four volunteers from the second year undergraduate psychology course, none of whom had encountered any work on syllogisms before. They were split into two groups, A and B, with a slightly different experimental procedure used in each case.

A standard computer terminal (a Freedom 100) was used for the experiment. This has a 10 inch (diagonal) green crt, which displays twenty four lines of eighty characters, and a detachable keyboard. The terminal was situated in a quiet, evenly-lit room, and the screen was angled to minimise glare. Subjects were shown there without any previous information about what the experiment was about. They knew only that the experiment would take about an hour, but that they should have the following half hour clear so that they would not have to rush towards the end. The terminal was connected to the departmental VAX 750 running Berkeley UNIX Version 4.1, configured in *cbreak* mode. This means that the characters typed by the subject were available to the experiment program at once, as opposed to only once the user had typed a carriage return as is usual on time-sharing systems.

Once comfortably seated, subjects began by reading through three screens of information about using the system, each of which told them to press a key to get the next page when they were ready. These screens warned subjects to be sure to tap the keys rather than press them, since the terminals automatically send a stream of characters if a key is held down. They also explained the mechanism for giving answers (which differed from group to group). Group A, who would be actually typing their answers, were told how to use the "back space" key to correct their errors, and it was pointed out that they could abbreviate as much as they liked, and that "most of the time all the words you might want to type will begin with different letters" so as to facilitate this. Group B, who would be composing their answers from "skeleton sentences" (see below), were given a practice item (the time of day) to ensure that they got the feel of selecting from the table.

Both groups were told that they would be timed, and that they should therefore be sure to decide what their reply was going to be before they started to give it. They were also told not to worry about typing speed, because the computer did not mind how long they took to type.

After this the subjects were shown a screen containing Text 1 (For Group B, the description of the mechanics of answering was suitably modified), which for the first time tells them what they would actually be required to do during the test. This is believed to be very close to an exact reconstruction of the instructions that Johnson-Laird and his co-workers use, with the obvious differences related to typing not writing answers, and the addition of the question concerning their confidence.

At the end of this screen the subjects were verbally asked if they had any queries or problems. Most subjects said they were confident that they knew what was required of them. Several asked for confirmation of their (correct) assessment of the polarity of the confidence scale, and were told that they would be reminded of this each time a confidence was requested. A few subjects asked if they could have pencil and paper, and were told that they could not because the experiment was attempting to see how people tackled the problems in their heads. None of the subjects asked (or were told) how many items there were in the test, so would not be tempted to pace themselves. The subjects were then verbally reminded about the importance of separating typing and thinking,

---

You are going to take part in an investigation of the way people combine information in order to draw conclusions from it. You will be given a series of pairs of statements about people you are to imagine assembled in a room. Type in what follows from each pair of statements about the occupants of the room. Base your answers solely upon what can be deduced with absolute certainty from the pairs of statements. Your conclusion must in one of the following forms :-

all thingies are doobries.
some thingies are doobries.
no thingies are doobries.
some thingies are not doobries.

For some pairs of sentences you may decide that there is no such conclusion which follows. In these cases you should just type a ".".

The examples will vary in difficulty, so you will be asked at the end of each item to type a number between 1 and 9 to show how confident you are of your answer.

Give your answers both as accurately and as quickly as possible, and remember to decide your answer before you start typing it.

When you have read this screen, say so and I will answer any queries.

Text 1. Final screen of subjects' instructions

---

not only because they were being timed, but also because re-thinking once they had started typing seemed to confuse people.

Once the subjects had started they were left alone in the room, and only looked in on once or twice during the test. Each subject was presented with one syllogistic problem in each of the 64 forms in a unique random order.

As mentioned above, the lexicalisation of the syllogisms was also unique to each subject, and random within a set of constraints. In Johnson-Laird's experiments it seems that problems were presented using single word terms, each word being used only once per set of syllogisms. This requires a vocabulary of nearly 200 single-word descriptions of people, all familiar words and none obviously carrying any strong connotations. Combining this with a stipulation that each item contain at most one term that the subject will (might) interpret as describing a person's employment (most people have only one job) leads to a requirement for a vocabulary of 128 other suitable descriptions of hobbyists etc.

Since Johnson-Laird's own vocabulary was unavailable, resolving the conflict between the need for so many suitable items and the desire to avoid unfamiliar words seemed a formidable undertaking. Even making the minimal change to allow multiple word noun-phrases (such as "butterfly collector") doesn't seem to generate enough variety (without resorting to "tennis player", "table-tennis player", "badminton player" etc.). At this point it is worth considering the motivation behind using different noun-phrases for each example, rather than having them all relating artists, beekeepers and chemists. Primarily it is an attempt to minimise the interference between items, although it probably also has the effect of making the task slightly less repetitive and featureless, and thus helping to keep the subject's attention. However, it seems likely that the exact details of the racquet sport mentioned in the example three items before is not going to have much effect on the amount of interference between items. For this reason, item repetition was tolerated, though with a minimum separation of two examples. Thus a vocabulary of 88 words was used, and no word was allowed to appear twice in any three consecutive syllogisms.

In addition to controlling repetition, every word in the vocabulary was marked as being in one or more of a range of categories, these being either *semantic* (e.g. "job", "sport") or *lexical* (e.g. "fans" or "collectors"). No two words in the same example were allowed to be members of the same category, nor start with the same letter. This was intended to ensure that the examples presented to a subject were as content-free as possible, and not too similar lexically. In addition, vocabulary items could be given a "social class", and all the words in each example were constrained to be in the same class (or at least, not to be definitely in different classes), so as to rule out socially implausible sentences such a "all of the bricklayers are ballet lovers". As a result of these measures, there seems to be no reason to suppose that any individual sentence employed would be any more distracting than those used in Johnson-Laird's experiments, but with the additional security that the effects of any bad word groupings would not accumulate.

Subjects in Group A were presented with the premises on the screen and prompted with the text "what, if anything, follows?". They responded by simply typing either their conclusions, or, if they thought there was no conclusion, simply "." on its own. They were then asked how sure

they were, and responded with a digit, at which point the program moved on and presented the next item.

Subjects in Group B were also presented with the premises, and asked the question "is there a sentence which always follows from these two?". They answered this by using the cursor keys to select either "yes" or "no" from a menu. If they selected "yes", the premises were removed from the screen and the subjects were presented with a multi-column menu, thus:

all		artists	are	artists
some	of the	beekeepers		beekeepers
none		chemists	are not	chemists

The predicates appear in the same order in column 2 and 4, but this is randomised for each item. The subjects then used the cursor keys to select one word from each column, thus composing a suitable conclusion, which appeared along the bottom of the screen as the words were selected. Once a complete conclusion was indicated, the premises were re-presented, together with the subject's suggested conclusion, and the cursor keys used to indicate certainty on an annotated scale from 1 to 9. If the subjects answered "no" to the initial ("Is there a conclusion") question, they were asked to rate their certainty on the same scale. At any time until they had given their confidence, subjects knew that they could return to the presentation of the premises to reconsider their answer. Subjects were timed until they finished looking at the premises for the last time.

It must be stressed that the method employed for Group B is **not** a multiple choice test, in that they are not presented with a list of possible conclusions. Nor can the menu readily be used as a memory jogger, to highlight the possible conclusions, since it cannot be seen simultaneously with the premises.

After all sixty four items were presented, the subjects were informally "debriefed", with the aid of a short questionnaire. This was intended primarily to check that the subjects had not had any problems with the computer and to pick up any clues to how (they thought) they had done tackled the task. Where it is used below, information gathered in this way is marked by (*Q*).

6.2.2. Results

The experiment was conducted both to obtain data on the responses of individual subjects and to give some idea of the repeatability of the experiment reported in Johnson-Laird and Steedman (1978). Discussion of the details of the responses of individual subjects will be deferred until after a new account of the processes involved has been presented, and the discussion in this section focuses mainly on the properties of the group distributions.

6.2.2.1. Categorisation of Responses

Even though the instructions told subjects to only offer conclusions in one of the four syllogistic forms, subjects in Group A (typing their answers) often offered other forms of conclusions, a problem also reported by Johnson-Laird and Steedman. Some of these could not be seen as in any way equivalent to any of the required response forms (e.g. "some artists may be beekeepers"). Others were synonymous with syllogistic forms (e.g. "some dog lovers don't jog"), or could be unambiguously recognised as more specific forms of them. Thus, in response to

the premises "all of the campers are traffic wardens, some of the traffic wardens are photographers", the conclusion "some of the traffic warden campers are photographers" can be taken as equivalent to "some of the campers are photographers". Another common error of this kind is illustrated by

some of the A are B
all of the B are C
therefore
some A are B and C

This undeniably states that "some A are C", though not in syllogistic form, and subjects were treated as having given this reply.

However, contrast this with the very similar, and common, deduction

Some of the A are B
All of the C are B
therefore
Some B are A and C

Now if some B are both A and C, then necessarily those same individuals can be described by saying "some A are C". This acceptable syllogistic form is a necessary consequence of the response given, and undeniably reflects the subject's grasp of the situations allowed within the problem. Therefore, one could argue that, as with the previous example, the subject should be treated as having given a response of this kind. However, there is a real sense in which explicating this is a deduction, and arguably comparable in complexity to the deduction that the subject is being asked to make. Therefore, and in contrast to the previous example, replies of this sort were deemed uncategorisable, examples of what Johnson-Laird terms "partially digested premises".

It was suggested in Section 6.1 that a change of this kind, either deliberate or unnoticed, in the way that such ill-formed responses were interpreted could well account for the differences between the results of the "valid syllogisms" experiment given in Wason and Johnson-Laird and those appearing in Johnson-Laird and Steedman.

Uncategorisable conclusions are clearly an undesirable waste (i.e. of subjects). They arise from the fact that although the subject has been told what is required, they have not fully grasped the details (or remembered the required forms). However, any attempt to improve subjects' understanding of the problem would almost certainly teach them more about the structure of syllogistic problems. As described in the previous chapter, Simpson and Johnson (1966) and Dickstein (1975) based their experiments on the fact that instruction can very quickly influence subjects' behaviour. Since the experiment is attempting to investigate how subjects "naturally" tackle syllogistic problems, such influence is obviously highly undesirable, and thus instruction to the subject must be minimised.

One of the benefits of the methodology employed by Group B is that subjects are very tightly constrained in the form of the conclusions they can offer, but without in any sense giving them any clues to the structure of the problem. Unfortunately, however, the situation is not quite perfect, and the menu-picking mechanism allows subjects to give problematical answers, of the form "all of the X are not Y". It is not obvious that this could be prevented without moving very close to a multiple choice format, a potentially crucial alteration of the problem. Three subjects actually did

offer conclusions of this awkward form, which is potentially ambiguous between "no X are Y" and "some X are not Y". They were contacted and asked both whether they considered the sentence ambiguous and what they thought they had meant by it. None had any doubt that the former was the intended reading.[3]

6.2.2.2. Rejection of Subjects

In an ideal world, subjects would fully understand what was required of them before commencing the experiment, and would simply do their best to achieve it. In the real world, people tend to forget or ignore instructions with depressing regularity. In the light of this, it is an unfortunate feature of the experiment that subjects are not given any feedback concerning the acceptability of their answers. As a result, any subject that does not fully grasp what is required may continue to offer unacceptable conclusions throughout the experiment. The problem is confounded because the experiment is intended to explore how people naturally go about solving verbal reasoning problems, and thus it is imperative that the instructions in the task do not suggest the way it should be solved.

Subjects' instructions asked them to offer a conclusion, in syllogistic form, that followed from the pair of premises. Most subjects realised that this meant that the conclusion they were to offer should need both premises to justify it, and thus should not include the middle term. Most others spontaneously realised this very early in the experiment (see below). Unfortunately, some subjects clearly did not grasp all the constraints placed upon their answers. In particular, one subject in Group A offered 24 "uncertain" conclusions, e.g. "some A may be C", while one subject in Group B offered 27 conclusions that included the middle term. Analysing subjects in terms of the number of such inappropriate responses given indicates that both these subjects, and another in Group A (20 bad answers), are more than two standard deviations from the mean of their respective populations (and more than three from the means of the populations left by removing them). If this is taken as indicating that these subjects have not grasped the problem, then they should be rejected from the group data. This leaves 10 subjects in Group A and 11 in group B, or a combined total of 21. Amazingly, none of the subjects rejected was the worst in their group in terms of valid conclusions given!

6.2.2.3. Distribution of Responses

The table of combined results from the two groups within the experiment is presented in Appendix A. The general distribution of responses is very similar to those obtained by Johnson-Laird and his co-workers, with one or two preferred conclusions for most problem types, although in some cases as many as 4 or 5 different conclusions were repeatedly offered. However, a number of specific features of these results are worthy of attention.

---

[3] For many (most?) people, the author included, there is no ambiguity at all. However, just after the the experiments were completed, an episode of an American detective serial was shown on television. In the very last line, the hero's colleague criticised his assumptions about the motives of an innocent female character, saying "you see, all women are not looking for a husband"!

6.2.2.4. Comparison with Johnson-Laird's Experiments

The most obvious difference between this experiment and those performed by Johnson-Laird and his co-workers is the method of example presentation. One can deduce from the report of the original experiment that examples were presented to the subjects on cards (or pieces of paper), on which they wrote their responses. Clearly, presentation using a crt and answering by typing are very different from these methods. Although some subjects said they were initially a little nervous of using a computer, all said (*Q*) that they quickly settled in and did not find it a problem. None of the subjects from Group A found (*Q*) the typing much of an effort, even those who had never done any typing before. Similarly, all the subjects in Group B reported that they were happy with picking words from the menus. No current theory suggests that these differences should have any effect, but this ought to be confirmed.

The experiment was intended to be an example of a paradigm established exclusively by Johnson-Laird, even though it deviates in a number of methodological features. In order to tell whether these innovations have introduced abnormal factors into subjects' behaviour, it is important to determine how well the results obtained agree with those that he obtains. The grossest characteristic of the results would seem to be subjects' overall accuracy: i.e. percentage of correct responses given. These, and the overall accuracy scores, in terms of the percentage of subjects responses that were valid, are shown in Table 6.1. These show that the new results fall close together, and are fairly central in the range of results that Johnson-Laird himself achieves. However, this is a very gross measure of the similarity of the experiments, and what is really required is a means for quantifying the agreement between batches of distributions of conclusions, although there does not seem to be a standard method for doing this. A suitable measure should be symmetrical, insensitive to the size of the subject populations but sensitive to the extent to which the two populations give the same responses for each problem form.

One such measure seems to be the area of overlap of the superposed normalised histograms of the responses for each of the 64 syllogistic problems. That is, for each syllogism form, the histogram of responses for each group of subjects is scaled to give an area of 1 (to eliminate the

Result Set	Valid Conclusions	Correct Responses
Steedman1:	54%	58%
Steedman2:	65%	69%
Bara:	36%	38%
Inder (Group A):	44%	49%
Inder (Group B):	53%	55%
Inder (Overall):	49%	52%

Table 6.1: Gross Properties of Group Responses.

effect of the size of the subject population), and the lower of the two frequencies for each possible response used to give a cumulative total. An alternative measure, obtained by summing the products of corresponding column heights, might also be considered reasonable.

In either case, there is a decision to be taken regarding uncategorisable answers. Clearly, they should not simply be placed in a "miscellaneous" category and allowed to count towards the similarity score, since they could well be different bad responses. However, it is debatable whether they should be included in the normalisation process, and this could be significant, since their presence will affect the "worth" of the remaining subjects. Consider an extreme situation where all the subjects in population X offer one particular response, while in population Y three-quarters of the subjects give that same response, and the other quarter give an uncategorisable reply. If the normalisation takes place over all subjects, the problem will score a similarity of 0.75, while if the populations are normalised over those subjects giving a sensible reply, the score will be 1.0. Which score really reflects the "similarity" of the two response populations? If a bad response is regarded as purely an execution error, then one could argue that the subject should be completely ignored, and only the sensible responses normalised (i.e. score = 1). However, if one considers that the tendency to induce a bad conclusion is a property of the syllogism form or subject population, then the normalisation should be over all the subjects in the group (i.e. score = 0.75). Both measures have been evaluated and found to be very similar, so henceforth only the latter option will be discussed.

Unfortunately, Johnson-Laird and Steedman present a subtle problem for any attempt to compare response distributions. They do not report all their subjects' responses, but only those that are given by two or more subjects, **or** that are predicted by their theory.

The motivation for rejecting the uncommon replies is, presumably, that since they are given by only one subject, they might well be execution errors. Omitting such extraneous responses from tables intended for "human consumption" might well make the key features easier to grasp (not to mention saving printing space). However, no such motivation can apply to the calculation of group statistics, where the unreported responses could obviously affect any measure of similarity. If any response has no corresponding response in the other data set, it will make no contribution to the similarity of the responses. In contrast, where both groups exhibit the same "aberrant" response, the shared tendency to exhibit it will contribute to their measured similarity, as indeed it seems it should. Obviously, this means that removing responses that are given only once will tend to reduce the similarity scores of the group, and doing so in a theory-specific manner will thus alter what is intended to be an objective property of pairs of response distributions. Ideally, all unique responses should be included in the calculation of any measure of the similarity of response distributions, but since in the case of Johnson-Laird and Steedman, these results are not reported, this is impossible. As a result, to ensure that all sets of data are evaluated on an even (and non theory-dependent) footing, any response given by only 1 subject should be ignored, whether predicted by any theory or not.

These metrics were used to evaluate the "distance" between each of the sets of results that are available. The different metrics gave the same ordering of the similarities, although they

differed in the relative numerical measures assigned to them. Table 6.2 shows the results of both measures (overlap and column product) for all pairs of sets of results, filtered to ignore all responses given only once. Normalising over actual responses, rather than possible responses, produced very little change in the relative similarities, as did using all possible results (as opposed to dropping those given once only). The following remarks hold of all these possible variations of the similarity measures.

The greatest similarity (75% overlap) is between the two sets of data presented by Johnson-Laird and Steedman, which involve the same subjects on different occasions. This fact becomes more significant in the light of the theoretical discussion below. The least similarity (55% overlap) is between data sets 3 and 2 in Table 6.2), which are the sets of results that form the focus of

Sets of results

1) Steedman1
2) Steedman2
3) Bara
4) Inder (Combined)
5) Inder (typing replies, 10 subjects)
6) Inder (menu picking, 11 subjects)

Column Overlap

	1	2	3	4	5	6
1)	—	75 (13)!	61 (17)	70 (12)	62 (15)	68 (13)
2)	75 (13)!	—	55 (19)	68 (13)	61 (17)	67 (13)
3)	61 (17)	55 (19)	—	62 (17)	57 (18)	58 (20)
4)	70 (12)	68 (13)	62 (17)	—	**	**
5)	62 (15)	61 (17)	57 (18)	**	—	62 (17)

Column Product

	1	2	3	4	5	6
1)	—	49 (17)!	37 (18)	41 (14)	38 (15)	42 (15)
2)	49 (17)!	—	37 (20)	44 (16)	42 (17)	46 (17)
3)	37 (18)	37 (20)	—	35 (16)	34 (16)	35 (18)
4)	41 (14)	44 (16)	35 (16)	—	**	**
5)	38 (15)	42 (17)	34 (16)	**	—	38 (16)

"!" denotes similarities between same subjects' first and second attempts
"**" denotes entires with subjects common to both groups

Bracketed figures show the standard deviation of individual syllogism forms.
Group responses were normalised over the size of the group.
Responses given only once were ignored.

Table 6.2: Mean Percentage Similarity of Distribution of Results.

Johnson-Laird's papers – i.e. Bara and Steedman2 (i.e. their 2nd "sitting"). In this context, it is clear that the new data are very much more like both sets of Johnson-Laird's results than these are like each other. Of course, one might argue that that one should consider the Steedman1 (i.e. the subjects' first sitting), since in both the other experiments the subjects were tackling the problems for the first time. If one does this then, contrary to the remark by Johnson-Laird and Steedman (1978, P371) that "the pattern of results is very similar in both tests", the similarity scores change considerably. The first sitting turns out to be even more like the Inder's results, and much more like Johnson-Laird and Bara's. These observations suggest that the new experiment can be seen as falling centrally within the paradigm that Johnson-Laird is seeking to establish.

Inder's experiment actually gives results for two groups of subjects. Both the subjects and the materials employed were very similar in the two groups, so one might well expect that they would produce very high similarity scores. However, this is clearly not the case – while the two groups do score highly, other pairs, under less closely similar circumstances (i.e. either of the two Johnson-Laird and Steedman sittings and the Inder's whole group), score even higher. This difference is partly attributable to the fact that the measures shown are based on data filtered to remove responses given only once: Inder's separate groups are small – approximately 10 subjects – and thus the effect of each response is emphasised. If all the similarities are re-computed using all available data, all values increase by about 2%, except that relating the Inder's two sub-populations, which increases by 8%, giving a value as high as any obtained using groups of different subjects, but no higher. This means that the differences between the two halves of the Inder's subjects (which show a 70% overlap) are as great as they are between completely independent populations.

Since the two groups could be expected to produce very similar results, it is tempting to ascribe this difference in performance between the two groups in the experiment to the factor that distinguished them: the mechanism for giving answers. One could argue that composing responses by selecting words from a menu influences the subjects' approach to the problem, and thus affects the conclusions that they offer. Unfortunately for this approach, the results of the menu-picking group are *very* similar to those obtained by Johnson-Laird and Steedman – much more so than the responses given by the subjects who typed their answers in full. This makes it hard to see how any explanation can ascribe the inter-group difference to a change of experimental method, since such changes would have to exclude, or cancel out in, the case of menu-picking.

This leads to the conclusion that the difference in the results reflects the genuine differences in the subject populations. However, the subjects employed were all members of one particular university course, and thus represent a very homogeneous population. This suggests that the observed dissimilarity or variability – a 70% overlap – reflects the minimum level of variation within even homogeneous populations, akin to the physicists' concepts of Brownian Motion or Zero Point energy. Moreover, since the similarity of Inder's results to Steedman2 are close to this level, this would indicate that they represent very similar populations. In contrast, something about the experiment conducted by Bara – either the subjects or the materials employed – is distinctly different.

6.2.2.5. Effects of Material

As described in Section 6.1, it is not obvious that it is possible to reliably identify (and thus control) the semantic content of sentences. This problem was circumvented by lexicalising each item differently for each subject, thus ensuring that every item is equally affected by any biases in the material. Every subject was asked (after the experiment) whether they had been confused or influenced by the actual words in the examples. Several subjects reported (*Q*) that, despite the semantic screening, they had found one or two examples which struck them as particularly silly,[4] although they also claimed (*Q*) that they had not been bothered by the anomalies that they had noticed. Since neither of the experiments reported by Johnson-Laird appear to control for this potential effect the deviation from exact replication in this matter would seem to be justifiable as improving experimental technique.

Given that results are available from more than one experiment, it is possible to try to demonstrate, and even quantify, the effect of "leading" material. With this objective, the distributions of responses were analysed by examining the differences in the sets of results obtained for individual syllogism forms, using the similarity measure described above. Under ideal conditions, the experiments being compared would be identical in every respect except for materials, and thus the effect of the material would be highlighted by the different distributions of results obtained.[5] Unfortunately, the actual experiments being considered differ not only in the materials employed, but also in the subject populations, which are drawn from different cultures. The response distributions different experiments yield for the same problem form differ in many ways: they never agree exactly, and the degree of similarity between them varies widely from item to item. This means that just because two experiments yield response distributions that are dissimilar – or even less similar than average for those two populations – it is not possible to take this as evidence of the effect of the material employed.

However, this problem can be partially surmounted, since the results from three experiments are available. If one group of subjects has been influenced by "leading" material, their response pattern will be less like the responses of **both** the other two groups, whereas the similarity of their results will be unaffected. Therefore, the results from the three experiments (Johnson-Laird and Steedman (first sitting), Johnson-Laird and Bara and Inder) were examined for the effects of anomalous material as follows. For each pair of sets of results, the similarity scores for each syllogism form were calculated and expressed in terms of standard deviations from the mean (for that pair of sets of results). Then, the "oddness" of each syllogism form was calculated as the difference between the best (most similar) two scores (i.e. the difference in similarity between the best score and the best of the other two). This is thus a measure of the extent to which one similarity is greater than both the other two, and thus reflects how well the situation fits that expected from anomalous material. The response distributions for the "oddest" syllogism forms

[4] The only specific misleading example quoted was the "guitarist/manager" example quoted in Section 6.1. This example was 63rd in the set of 64, and thus was recently encountered and fresh in the subject's mind.

[5] In fact, the differences observed between Inder's two, very similar, groups suggest that this ideal may be unobtainable.

are shown in Table 6.3.

In the case of the **1OO** and **1II** syllogisms, the great majority of Bara's subjects offer particular conclusions, while most subjects in the other experiments recognise that there is no valid conclusion. Similarly, for the **4OA** syllogism Bara found that the most common response was a particular positive conclusion, which was rare among the other groups of subjects, who generally offered particular **negative** conclusions. Thus it could be argued that in these three cases Bara's experimental materials have led his subjects to offer conclusions that, on the evidence of the other two groups, they would not normally have done.

The situation for the **1OA** syllogism is less clear cut. Here, it is the Johnson-Laird and Steedman subjects that are unusual, and much of the response difference concerns the "Gricean"

Oddness	Problem	Response	Frequency			
			Inder	Steedman		Bara
				1st	2nd	
			(21 Ss)	(20 Ss)		(20 Ss)
3.1	**1OO**	**NVC**	15	14	19	—
		O(a, c)	2	4	0	15
2.5	**1II**	**NVC**	12	13	16	2
		I(a, c)	7	7	3	18
2.0	**4OA**	**NVC**	2	3	5	4
		O(c, a)	12	12	14	—
		I(a, c)	1	3	0	—
		I(c, a)	3	—	—	9
1.9	**1OA**	**NVC**	1	12	9	1
		O(a, c)	8	5	9	9
		I(a, c)	7	—	—	4
1.7	**1EE**	**NVC**	17	9	13	10
		E(a, c)	3	10	5	11
1.5	**200**	**NVC**	13	17	16	10
		O(c, a)	4	0	3	6
		I(c, a)	2	—	—	2
1.3	**2OI**	**NVC**	12	9	16	4
		O(c, a)	4	8	3	13

"Oddness" rating is explained in the text.
— indicates no figure published, but fewer than 2.

Table 6.3: Problems Where Response Distributions Differ Widely.

responses, which they never gave.[6] On this particular item, the majority of them recognise that there is no valid conclusion, whereas only one subject from each of the other experiments does. This might suggest that the material (or more accurately one of the two sets) employed for this problem allowed subjects to recognise the invalidity of the otherwise-favoured **O** conclusion. This is supported by the fact that on the second sitting, when the subjects were undeniably more skilled in solving syllogistic problems, their performance gets **worse**. It seems this can best be explained by suggesting that not only are the subjects more skilled at solving syllogisms (so their performance should improve), but they are also less prone to be influenced by the distracting material. Precisely similar arguments hold for the **2OO** syllogism, although the differences are smaller.

In the case of the **1EE** syllogism, the difference arises from the fact that very few of Inder's subjects suggest the otherwise common **E** conclusion. This cannot be put down to Inder's use of misleading material, since no two subjects used the same lexical items. However, a possible alternative explanation is suggested by the fact that the Steedman2 responses are much more like those of Inder. By calling once again upon their increased resistance to the effects of misleading material, it is possible to suggest that **both** Steedman and Bara used misleading materials on this particular problem form – i.e. material that biased their subjects towards a negative conclusion. This is not particularly surprising since, as mentioned above, all of Bara's conclusions were intended to involved two professions, and thus might be expected to provoke a negative bias. The **2OI** syllogism illustrates precisely similar trends.

The above analysis has suggested that there is evidence of the effects of material to be found within the differences between the three experiments examined, and therefore that the methodology of randomising the materials used on a per-subject basis is sensible. However the very fact that the comparison could be carried out at all suggests that for the most part the materials used are good enough – that is, most subjects are able to ignore the semantic biases they introduce, at least most of the time.

Finally, it is worth noting that every subject, including those rejected for not grasping the details of the problem, understood and was able to answer the interview question regarding misleading material. This is significant, since in order to understand this question it is necessary to realise that a syllogism can be solved independently of its actual content – that there **is** a correct answer with which the material can be incongruous. The fact that all subjects recognised this argues against any attempt to ascribe their reasoning errors to a "failure to accept the logical task", as Henle (1962) might suggest.

Given these observations, there would seem to be no reason to expect that the change of method ought to affect the subjects' performance on the task, and certainly there is nothing in any of the "standard" theories of syllogistic reasoning that would be at all affected. It seems, therefore, that it is reasonable to look on this experiment as a replication of those reported by Johnson-Laird (1978, 1983).

---

[6] These were mentioned in Section 6.1 above. Notice that this is the only "Gricean" influence discernible in the six "oddest" problem forms, which suggests that this global response difference has been effectively screened out.

6.2.2.6. Effects of Figure

There is clear support for the Figural Effect, which is observable in each individual subject (although one is marginal). The assessment of the relative "easiness" of the figures is less clear-cut. In terms of the number of correct responses given per subject, the figures are very similar, with no discernible trend. However, Johnson-Laird prefers to work with the number of valid conclusions offered (i.e. a correct reply that there is "no valid conclusion" is not counted). In this measure Inder's results once again support those of Johnson-Laird (See Table 6.4).

6.2.2.7. Timings

The timing of subjects' responses is another area where Inder's method differs from those reported by Johnson-Laird. Johnson-Laird's subjects were (presumably) timed by the experimenter, using a stop-watch, started when the subject saw the example and stopped when the subject started to write a solution, or in some other way showed that they had reached a decision. The current experiment attempts to implement the equivalent timing, by noting the delay between example presentation and first keystroke of solution. Many subjects expressed the opinion (*Q*) that the computer method was far less obtrusive. Even though they were aware that they were being timed (having been both told and verbally reminded), they all said that they thought they were much more relaxed and able to concentrate on the problems than they would have had they known that someone was watching over them with a stopwatch.

Valid Conclusions

Figure	Bara	Steedman		Inder
		I	II	
1	49	63	70	55
2	46	53	60	54
3	33	53	60	39
4	22	49	69	49
Overall	36	54	65	49

Correct Responses

Figure	Bara	Steedman		Inder
		I	II	
1	28	56	68	51
2	38	53	62	51
3	44	64	69	51
4	42	60	76	56
Overall	38	58	69	52

Table 6.4: Effect of Figure on Percentage of Correct Responses and Valid Conclusions Offered.

Because the program was running on a time-shared system, the response time of the system could not be guaranteed, and this does open up the possibility of errors in the timing data. However, extensive use of the experiment program under system load conditions similar to those at the time of the experiment suggests that the system usually (effectively always) responds within one second, and always within two. Further, the program always does its timing before beginning to respond to the users key-stroke. Since no user (*Q*) experienced any perceptible delays, this would again indicate that no timings were upset by system response factors by a more than a second. This interval is small compared to the typical response times observed, so while the fine details of the timing data are suspect (individual data could be out by a second or just possibly two), the fact that such errors would be randomly distributed means that there is no reason to suppose that the general trends shown in the timing data should be particularly suspect.

The average response times for specific problem forms vary between 19 and 71 seconds, although as mentioned above, inaccuracies in the method of timing might be responsible for raising these figures by one or two seconds. Individual responses varied between 5 seconds and 235 seconds (i.e. over four minutes). The mean response time is 40.8 seconds, the median is 30 seconds and the first and third quartile times are at 17.7 and 50.3 seconds respectively. This indicates that the response is skewed, with most responses between 10 and 60 seconds, but with some very long pauses for thought: the 8th and 9th deciles are 59.4 and 83.5 seconds respectively.

Although not considered a particularly important feature of the experiment, subjects' response times stand in stark contrast to those reported in Johnson-Laird and Steedman (1978). All the times recorded are much slower than theirs, by an amount that cannot be attributed to the limited accuracy of the timing information. Some of this difference can be put down to the fact that subjects felt more relaxed and free to think about the problem in hand when being tested by a machine. Even so, this is unlikely to explain the size of the difference, which is an order of magnitude in some cases. However, the longer times do seem to be considerably more realistic. They are also comparable to those obtained by other experiments using computer-driven apparatus. One of Johnson-Laird's undergraduate students at Cambridge reports average response times between 17 and 35 seconds (Levy, 1984), while Frase (1968) reports conclusion **validation** times averaging 26 seconds.[7] It is hard to believe that a naive subject can read the premises and recognise the middle term, let alone solve the syllogism, as fast as some of Johnson-Laird and Steedman's reported times (down to 1.9 seconds **average**) seem to require.

[7] It is possible to speculate that this timing difference – Frase's 26 seconds for validation versus Inder's 41 seconds for solution – may reflect the difference in task. It could be explained by suggesting that, unlike those validating a given conclusion, subjects trying to find a conclusion usually consider more than one possibility. However, much of the work in falsifying one conclusion is relevant to the evaluation of another (i.e. the same situations must be considered). This means that explanations based on considering multiple conclusions must either postulate that many conclusions are tried, which is problematical since valid conclusions are often overlooked, or that subjects cannot retain their "intermediate working" between considering putative conclusions. Alternatively, one can assume – as does the new theory to presented in Chapter 7 – that the task of actually generating the candidate conclusions is itself time consuming.

6.2.2.8. Confidences

The experiment attempted to assess the subject's confidences in their responses on a scale from 1 (labelled "completely unsure; almost a guess") to 9 (labeled "absolutely certain; no question of error"). The results gathered have not been statistically analysed, but several notable features are immediately apparent. All subjects offered a range of responses, although (as is usual with such measures) some subjects clustered their responses into only the "confident" end of the scale. Almost all showed some evidence of guessing. Several simply gave some responses a confidence rating of 1. Most produced a clearly bimodal distribution of confidence values, with a small group of responses being separated from even the tail of the main spread of responses by at least one completely unused response category.

Moreover, there is some evidence to support for a suggestion, raised at an Edinburgh workshop by Terry Myers, that that subjects are using the "no valid conclusion" response as a "dustbin" category. Although for the most part these responses are supported as confidently as any other, in about half the subjects the responses categorised as guesses are overwhelmingly or entirely of this form. What is more, these low-confidence responses are by no means hurried: often they are particularly slow. This suggests very strongly that the subject **recognises** that the problem has not been properly dealt with, and offers the "no valid conclusion" response because they have to say something!

6.2.2.9. Individual Results

One of the central motivations for Inder's experiment was to obtain information on the behaviour of individual subjects, and the way this changed with practice. Four subjects performed a complete set of syllogisms once a day for four or five consecutive days. The details of the responses of these subjects will be discussed below. However, the following general properties can be observed.

(1) All the subjects got appreciably faster, with the time take for the total test typically falling from about an hour at the first sitting to about 25 minutes for the last. In terms of individual examples, the mean time decreased from 43 seconds to 20. The average item time for each individual subject is shown in Table 6.5.

(2) Subjects got more accurate, in terms of the number of valid responses they offered, with average scores rising from 42 on the first sitting to 48 on the last. This improvement was observable in three of the four individual subjects, and occurred in the absence of any kind of feedback or instruction.[8]

(3) Within a subject, responses from session to session converged, even though different materials and presentation orders were used on each occasion. Table 6.5 shows the differences in each subject's responses between the first and final pairs of results (i.e. first and second attempts at the test, and penultimate and final time). In two cases (incidentally, the two males) this convergence is obvious from the number of items that drew the same

[8] Subjects were asked not to practice between tests, and all said that they had not. There is no reason to suspect that this is not the case.

conclusion each time, and indeed one of them produced arguably identical responses on his 4th and 5th attempts at the experiments.[9] In the case of the other two subjects, the convergence is not obvious unless the differences between sets of results are sub-classified as shown. Those classified as "order" involve only the the order of the terms in logically reversible conclusions offered in symmetrical figures. This is a situation in which there is no logical basis for the decision, and the effect of incidental factors is minimised, and as a result the decision can be expected to be as close to random as any in the process. Those classified as "O or No" all involve consistently denying in one session that there is any valid conclusion to items which in the other session drew an O response. While this may appear an apparently arbitrary type of response to single out, it accounts for almost one quarter of the responses of one subject, and three-quarters of the variation between her two last sessions. Moreover, the discussion in Chapter 7 will suggest that this can be attributed to a single and common change in approach to the most difficult class of problems.

6.2.2.10. Introspections

The instructions for Inder's experiment deliberately included a suspect feature of Johnson-Laird's instructions: subject were told that they would be given problems about people "they were to imagine assembled in a room". At the end of the experiment they were asked what they remembered of the instructions (only two offered this feature, though all claimed it was familiar when they were reminded) and whether they had, in fact imagined people in any way. Only three

First Session

Subject	Mean Time per Item	Difference from Second Session			
		Same	O or No	Order	Different
1	29	31	8	3	22
2	53	48	3	2	11
3	67	47	4	1	12
4	22	45	4	3	12

Last Session

Subject	Mean Time per Item	Difference from Previous Session			
		Same	O or No	Order	Different
1	20	42	7	4	10
2	22	60	1	2	1
3	24	62	1	1	0
4	11	43	15	1	5

Table 6.5: Changes in Performance of Individual Subjects

[9] This subject – Subject 3 here – is discussed further as Subject C in Section 7.4.3. Note that since his answers were only 66% accurate, it cannot be argued that he was simply demonstrating his underlying logical competence.

subjects said (*Q*) they had, and they all said they soon gave up doing so because it was inefficient or because, as one subject put it, the imagined people "got in the way". Thus it seems that this feature of the instructions did not have any (conscious) "leading" effect on the subjects.

Subjects were also asked how they thought they had done the problems. The great majority of responses involved descriptions of using circles to represent sets. However, no subject used either of the terms "Euler circle" or "Venn Diagram". Furthermore, no subject was able to give any kind of succinct explanation of how they had employed the circles, although several attempted to describe complex systems that they claimed (plausibly!) to have developed during the experiment. While subjects' introspections are by no means conclusive, they do at the very least provide another datum to be explained.

6.3. Johnson-Laird's Theories

Johnson-Laird's theorising concerning syllogistic reasoning is largely shaped by his commitment to the idea of Mental Models. As described in Chapter 3, he argues on the basis of his experiments with Mani and Ehrlich for what he terms *physical* mental models, which exhibit similar properties and restrictions to the information structure argued for in Chapter 2. However, he recognises that a viable psychological theory must account for the fact that people can deal with more abstract information. As a result, he attempts to illustrate the relevance of his ideas by applying them to the specific task of syllogistic reasoning, recognising that to do this he must describe both the important features of the models involved and the processes that access them. However, for reasons that will be discussed in shortly, he believes that the kind of mental models that are suitable for physical situations are inadequate to deal with more abstract matters. This forces him to propose the use of *conceptual* mental models, which have significantly more representational power.

Although his ideas have undergone a major revision, there is much that remains constant throughout, and this will be presented first, before the details and differences are discussed. Johnson-Laird traces the inspiration of much of his theorising to a chance remark by one of the subjects in the first ("Wason") experiment.

> When he was asked to describe how he performed the task, he replied, referring to a specific premise, "I thought of all the little [sic] artists in the room and imagined that they all had beekeeper's hats on". This remark provided the germ of an idea for a new hypothesis about the semantic representation of quantified assertions: **A class is represented simply by thinking of an arbitrary number of its exemplars.**
>
> (Johnson-Laird and Steedman, 1980. P76)

The heart of the method that this inspired is most effectively communicated by (a loose reconstruction of) the explanation in (Johnson-Laird, 1983) of how one could go about solving a syllogistic problem using a room full of actors. Suppose that the premises were

> all artists are beekeepers
> some beekeepers are chemists.

One could begin by asking some of the actors to take on the roles of artists in some manner, such as clutching paintbrushes and a palette. Then one would ask them to also take on the role of beekeeper, equipping them with suitable headgear or whatever. Since there might also be some

beekeepers who are not artists, one could also similarly equip one or two of the remaining, "unartistic", actors. At this point, the actors in some sense "represent" the first premise, which Johnson-Laird represents as Fig. 6.1. The situation may be represented more compactly using a notation introduced in (Johnson-Laird, 1983), as in Fig. 6.2(i), where, following Johnson-Laird, a minimal number of actors are used to capture the situation, and the brackets round the solitary "b" indicate the uncertainty whether such an individual exists.

The second premise could now be dealt with by issuing some of the beekeepers, maybe some of those with paint brushes, with test tubes, thus allowing them to adopt the role of chemist as well. This would be reflected by a diagram such as Fig. 6.2(ii). Anyone studying this situation might be inclined to suggest that "some of the artists are chemists", which is true in the current situation, is a valid inference. However, this conclusion can be tested by getting the actors playing chemists to pass the test-tubes to other beekeepers, and in particular to those who are not artists. Doing this we can get a situation which corresponds to Fig. 6.2(iii), from which we can clearly see that this potential conclusion need not be the case, and in fact the premises remain true even if none of the artists are also chemists. Since some, none or indeed all of the artists may also be chemists, there is no relation that must hold and hence no valid conclusion that can be drawn.

In the light of this intuitive description, Johnson-Laird's actual theory (both before and after its revision) is apparently very straightforward. When confronted with a pair of syllogistic premises,

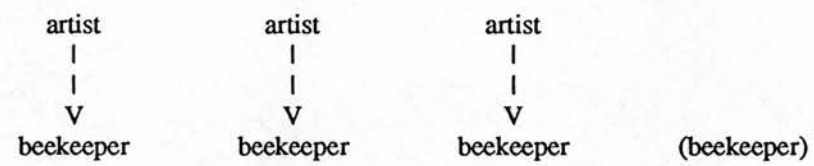


Fig. 6.1: A "Tableau of Actors" representing "all artists are beekeepers".
Notation from Johnson-Laird (Pre 1980).

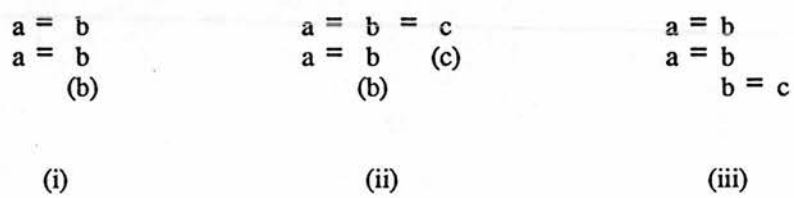


Fig. 6.2: Solving the **1AI** syllogism.
Notation from Johnson-Laird (post 1984).

subjects translate each premise into a simple mental model, and them combine these to obtain a representation of the complete problem. Johnson-Laird represents these models using notations like that used in Fig. 6.2, though he stresses that the actual number of entities of each type is arbitrary and may be changed later. From this model, the subject attempts to "read off" a conclusion, which is usually the strongest statement that can definitely be made about the situation represented by the model. If they cannot find such a statement, they know that the problem has no valid conclusion. Otherwise they may attempt to disprove the possible conclusion they have read off.

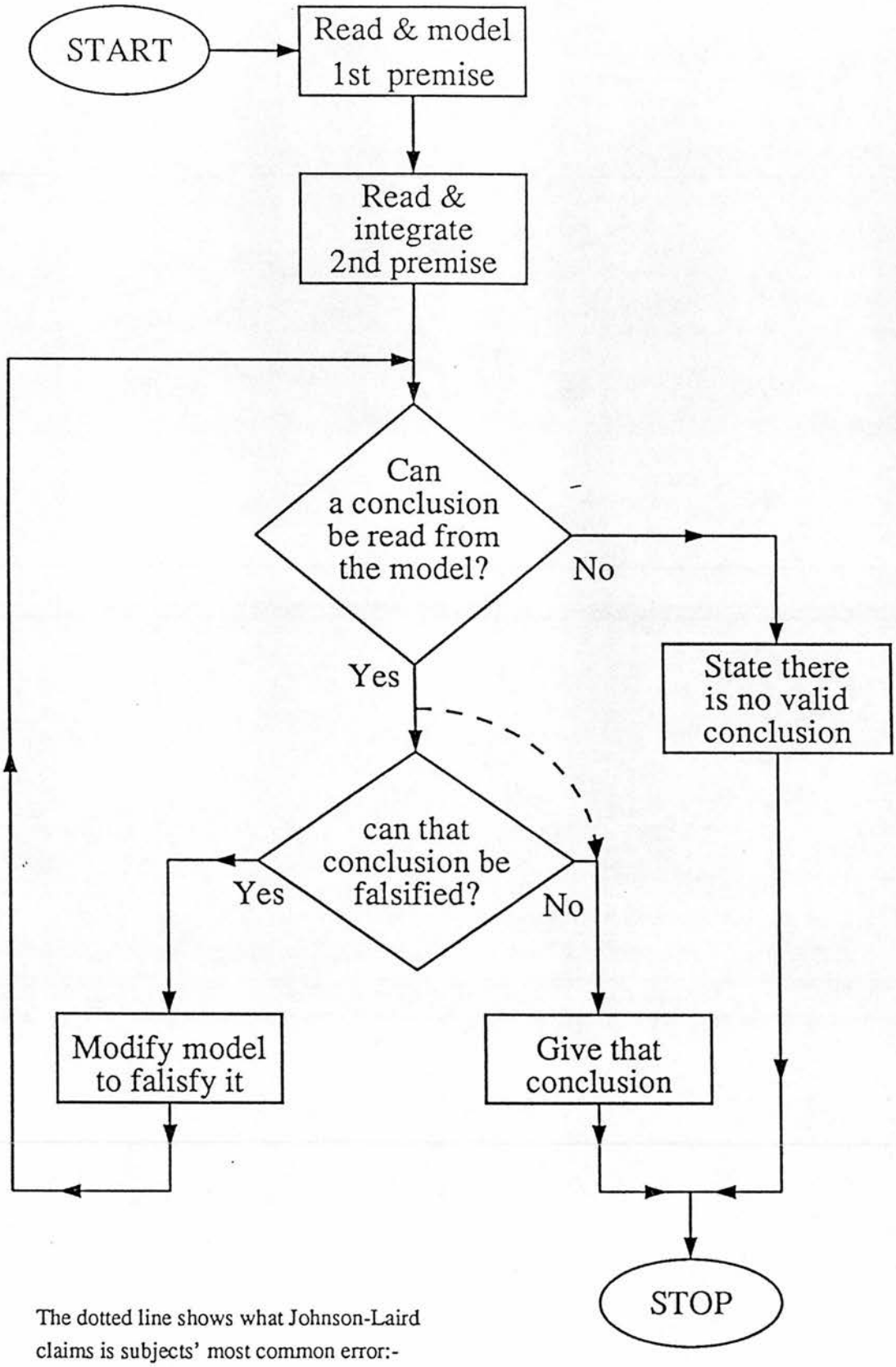
In order to disprove a potential conclusion, subjects will consider a range of modified models, generated with the intention of falsifying it. If they find a model which constitutes a counter-example to the conclusion (i.e. one in which the premises are true but the conclusion false), it is thus shown to be invalid and is rejected. In this case the subject will attempt to read another conclusion from the models that is consistent with all the situations considered, and once again consider testing it. Otherwise, if the model cannot be altered to defeat the conclusion, it is accepted as valid. The overall process is illustrated by the flow diagram in Fig. 6.3. If this modification process is repeated correctly, subjects will produce correct answers for rational reasons. Thus this theory meets the requirement for explaining how people can be (potentially) rational, though (crucially for Johnson-Laird) without employing any kind of deductive reasoning – that is, with no explicit combination of individual propositions in line with any kind of inference rules.

However, experiments show that subjects are far from perfectly logical, so to meet the criterion of empirical accuracy, Johnson-Laird must explain their aberrations. He does this by postulating that subjects often do not take the read/falsify process to completion. Instead, they break off prematurely, in the manner suggested by the dotted line in Fig. 6.3, and are thus led to accept a conclusion that could be falsified. In the example above this would lead to subjects being inclined to accept the conclusion suggested by the original model (namely, "some artists are chemists"), and results shows that this is in fact the most common response. One of the strong points of the theory is that the conclusions that can be read off the "intermediate" models correlate well with the observed responses of experimental subjects. They may also overlook a conclusion by failing to "scan the model in both directions" – that is, consider both term orderings.

Given the basic approach as presented so far, which has remained unaltered from 1975, one can appreciate the details of the two versions of the theory, which essentially determine the responses it predicts and which have been significantly revised between 1978 and 1984.

6.3.1. Mental Models and Syllogisms, 1975 — 1978.

One of the key features of the early mental models is that they employ directional links. Thus the models built in response to the **1AI** syllogism would be represented as shown in Fig. 6.4 (cf. Fig. 6.2). Johnson-Laird and Steedman (1978, P77) describe Fig. 6.1 as depicting a mental model representing "an arbitrary number of artists tagged as beekeepers". They go on to emphasise the directionality of these links, referring to "pointers" in memory and introducing the analogy of a list structure. They then go on to say that



The dotted line shows what Johnson-Laird claims is subjects' most common error:- omitting the test for falsifiability.

Fig. 6.3: Johnson-Laird's proposed syllogism solving process.

> The representation of each artist has stored with it the address of the corresponding representation of the [sic] beekeeper, and it is consequently easy to move from artist to beekeeper, but the representation of a beekeeper has no concomitant address of an artist, and the only way to move from beekeeper to artist is to search through all the artists until an appropriate link is found that leads back to the starting place.
>
> (Johnson-Laird and Steedman, 1978. P77)

This difficulty in following arrows "backwards" is responsible for the Figural Effect. Conclusions where the terms are encountered in the order indicated by the arrows are much easier to generate, and thus are produced more often. Within a model, the direction of any particular arrow is determined by the order of the terms in the sentence that caused the arrow to be constructed. Thus in Fig. 6.4, the arrows link the beekeepers to the chemists because the premise was "some beekeepers are chemists". Clearly, since the order of the terms determines the figures of a premise pair, the arrangement of arrows will vary between figures, and Fig. 6.5 illustrates (the diagrams corresponding to) the models initially built for each of the **AI** syllogisms. This shows that in the case of the **1AI** syllogism, a conclusion linking As to Cs will be following the arrows at each stage, and will thus be facilitated, while for the **2AI**, it is the conclusions linking Cs to As that are favoured. For the **3AI** and **4AI** syllogisms, A - C and C - A conclusions have to traverse equal numbers of arrows in the "wrong" direction, and thus there is no bias. This is precisely the distribution of responses that constitute the figural effect.

Within the early account of mental models, conclusion testing is presented as model modification by making or breaking links (i.e. a sequence of models is considered). Thus as Fig. 6.4 shows, an **I** conclusion to the **1AI** syllogism is disproved by breaking the bottom link in Fig. 6.4(ii), which falsifies the second premise, and correcting this by making the bottom link in Fig. 6.4(iii). Notice that another bracketed "b" has been introduced, which illustrates how the number of entities in the model can be changed.

Another feature of the early mental models is the representation of negative information by means of "blocked" links. Thus, the initial model built for the **3IE** premise pair is shown in Fig. 6.6(i), where the blocked arrows show that each C "isn't a" B. The mechanism for reading off a conclusion is sensitive to the presence of a blocked arrow and will thus produce negative conclusions. In this particular model, no conclusion can be drawn without going "against" an arrow, and hence no figural bias is predicted, so conclusions may be offered with either term order. In considering conclusions relating Cs to As, the mechanism may well produce the conclusion "no C are A", while in the opposite direction "some A are not C" can be read off.

"No C are A" is not, in fact, a valid conclusion, and Fig. 6.6(ii) illustrates how Johnson-Laird and Steedman suggest the model can be modified to defeat it: a link is made from each C to an A. Notice, however, that the leftmost C cannot be linked to the leftmost A, because this would lead to an incoherent model: a C which is linked to a B both by a chain of (transitive) identity links, and by a "negative" link. However, as Fig. 6.6(ii) illustrates, this is not a problem since links are not constrained to remain in their own column, but can link any pair of entities.

Well over 90% of the responses given by subjects in the experiment that Johnson-Laird and Steedman present were in categories predicted by their theory, which is clearly a strong point in its favour. However, this alone is not enough, since close to 100% accuracy is achieved by the

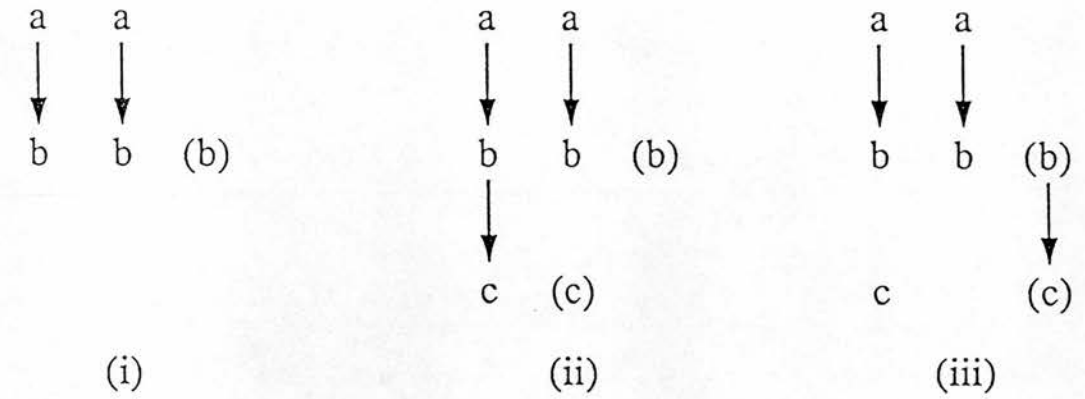


Fig. 6.4: Solving the 1AI syllogism.
Notation from Johnson-Laird and Steedman.

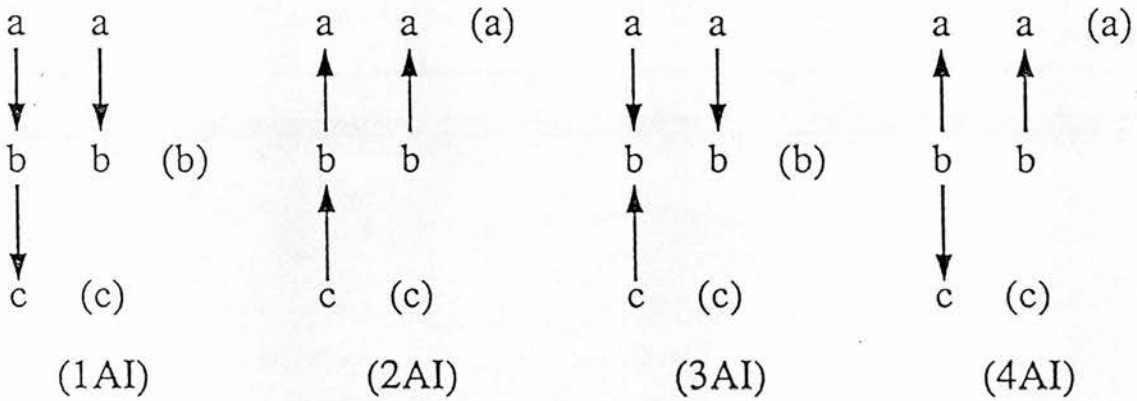


Figure 6.5: Initial models for the AI syllogisms.
Notation from Johnson-Laird and Steedman.

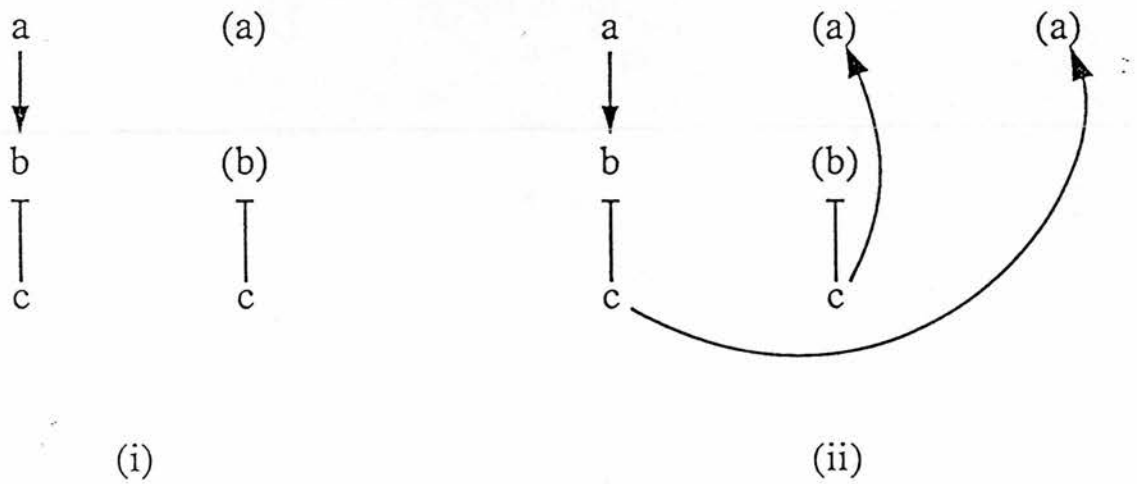


Figure 6.6: Solving the 3IE syllogism.
Notation from Johnson-Laird and Steedman.

vacuous theory that suggests that subjects responses will be in one of the 9 possible categories (four quantifiers with two premise orderings, and "no conclusion"). In addition, therefore, one must consider the proportion of categories the theory predicts will be given that actually **are** observed. Johnson-Laird and Steedman's claim that of the 213 categories (out of 576) that they sanction only 23 (11%) are not observed, and of these 16 were predicted to be rare anyway. However, if one ignores completely responses given only once, which is the policy Johnson-Laird and Steedman lean towards when presenting their results, the number of response categories predicted but not observed rises to 46, of which only half are predicted to be rare. This means that even in two sittings, 22% of the categories that are predicted to be observed are not given.

Johnson-Laird and Steedman recognise the importance of the applicability of their theory beyond syllogistic reasoning, and try to give some idea how mental models can be used for other types of sentences and deductions. They explicitly suggest that the entities in the models can be used to represent, and thus reason about, specific individuals (e.g. "Arthur is a Briton"), although they do not indicate how or whether model entities denoting specific individuals differ from those denoting "arbitrary" individuals in syllogism problems. Similarly, diagrams are offered to show how a model could be used to capture multiply quantified sentences and "vague" quantifiers (e.g. "most") and, by using different links, other verbs (e.g. "all the boys kissed some of the girls").

6.3.2. Problems with the Early Models

6.3.2.1. What do the Links Denote?

Johnson-Laird stresses the fact that, unlike propositional representations, mental models are isomorphic with the situation that they represent. Throughout both the quotations given at the start of the previous section, Johnson-Laird and Steedman refer to the links as interconnecting individual artists and beekeepers. This strongly suggests that the letters in these diagrams are to be taken as depicting (some kind of locus for information about) specific individuals, with the arrows depicting identity links. This is in line with the idea of mental models as sharing structure with the part of the world they represent, and also with the rationalisation of Johnson-Laird's ideas suggested by Lee (1984a).

Unfortunately, this is at odds with Johnson-Laird and Steedman's own comment (also P77) that "The arrows stand for the semantic relation of class membership (each artist *is a* beekeeper)". Except when one is explicitly doing mathematics, class membership is a relation between an individual and a *set* or *class*, whereas identity is a relation between one (description of an) individual and another. The confusion is deepened by the negative links in the early models: What, precisely, is represented in Fig. 6.4(i). It is tempting to say that the leftmost column represents an artist, tagged as a beekeeper, and also tagged as definitely distinct from a chemist. But this is **not** what it represents. The negative link is not a negative identity relation, but a negative **class inclusion** relation. Johnson-Laird and Steedman (P.79) describe a blocked arrow from an A to a B as "a negative which prohibits a positive link between the A and **any** B" (emphasis added). So the leftmost column of Fig. 6.4(i) must be glossed as representing both an artist tagged as a beekeeper, **and** a chemist which is not related to **any** beekeeper.

6.3.2.2. "No Valid Conclusion" Responses

Within the theory proposed by Johnson-Laird and Steedman, subjects realise that a "no valid conclusion" response is appropriate by analysing a range of possible situations and recognising that there is no (interesting) conclusion that is consistent with all of them. This is the only way that such a response can arise, which is a feature of that theory that is often found counter-intuitive. Terry Myers has argued (School of Epistemics workshop) that, in contrast to always being the ultimate result of careful analysis, such responses were often used as a "dustbin" category, being given whenever the subject felt the problem was insoluble or intractable. While the evidence from subjects' confidence in their responses (outlined in Section 6.2.2.8) is not conclusive, it certainly lends credence to Myers' suggestion, or at least seriously undermines Johnson-Laird's.

However, a more serious problem for the "result of deduction" view comes from Bara's experiment, where subjects were obliged to respond within 10 seconds. The great rise in the number of "no valid conclusion" conclusions offered clearly cannot be because a challenging time limit makes more subjects complete their testing, and thus suggests that at least some other mechanism must be able to generate the response. This is the crux of the argument that Johnson-Laird uses to motivate the changes from the 1978 theory.

6.4. Mental Models and Syllogisms, Post 1980.

Since his paper with Steedman, Johnson-Laird has considerably revised his ideas on the application of mental models to syllogistic reasoning, with slightly differing versions of his new ideas being presented in (Johnson-Laird, 1983) and (Johnson-Laird and Bara, 1984).

The most obvious difference between the theories is the least significant: the basic arrangement of the model diagrams has been rotated through 90 degrees, so that identity links now run horizontally (as in Fig. 6.7). In addition, negative information is encoded not by "isn't a" links, the problems with the interpretation of which were indicated above, but by the presence of a "barrier", with the sensitivities of the conclusion drawing mechanism suitably adjusted. Fig. 6.7(i) shows the model that would be built in response to the **1EE** syllogism, i.e.

> none of the artists are beekeepers
> none of the beekeepers are chemists.

As with the early theory, the initial model suggests a conclusion – "no artists are chemists" – which is not valid, but is still a frequently-drawn conclusion. Once again there is a stage of model modification, now taking the form of moving "properties" between entities, and Fig. 6.7(ii) depicts the modified model that Johnson-Laird and Bara suggest defeats the initial conclusion for the **1EE** syllogism. As in the earlier models, explicit negative information, now represented by the barrier, is distinct from uncertainty, as shown by the absence of a positive connection. Thus this model does not allow one to say that any artists are chemists: i.e. no conclusion can be read from it. In slight contrast, Johnson-Laird (1983, P113) does have positive links above the barrier, but conclusion can only be drawn if they are "common to all interpretations".

Another significant difference from the early version of the theory is the origin of the figural effect. The individual links between features are no longer seen as directional, but nonetheless there is a kind of directionality to the internal structure of the models, in that conclusions are read

Fig. 6.7: Models for the 1EE syllogism.
Notation from Johnson-Laird (post 1984)

off models from left to right. This ordering of items in memory is justified on the grounds that human memory can sometimes be seen as a kind of "first in, first out" mechanism. Johnson-Laird and Bara (1984, P31-32) point out that it is much easier to regurgitate a list of digits in the order they are presented than in the reverse order, and the same restriction can be illustrated by attempting to sing a verse of a favourite song backwards. However, these observations pertain to the ordering of information concerning independently perceived objects, whereas Johnson-Laid and Bara are attempting to apply the same restriction to the ordering of the properties of a single entity. Such an argument tempts one to ask awkward questions about, for instance, the ordering of the properties involved in someone's model of their mother's face, and the sequence in which they are relevant to recognising her.

As with the older theory, the order of the terms in the premises dictates the directionality of the models of the individual premises. However, in this version these premise models can only be combined to form a model of the problem if they are "congruent" – which is to say that the left-hand feature in the second premise model must match the right-hand feature in the first model. In other words, only models constructed in response to premise pairs in the first figure can be combined directly. In all other figures, additional operations must be carried out to manipulate the models into the required format. This involves either re-ordering the premises, or reversing the individual premise models, thus changing

a	=	b		b	=	a
a	=	b	into	b	=	a
	(b)			(b)		

This is highly reminiscent of conversion, and indeed the whole procedure of manipulating the situation into a specific figure is very similar to the approach of Chapman and Chapman (1959). They suggested that what they termed *probabilistic reasoning* could occur only in one figure, and could only be applied to the others after their premises had been converted, possibly illicitly. However, what Johnson-Laird proposes is not to be seen as a logical or semantic operation at all: Whatever interpretation is given to them, in terms of entities and their properties, these two

diagrams will be identical. Nonetheless, they are significantly different models for the purposes of Johnson-Laird's theories.

When examined in the light of Johnson-Laird's diagrams of mental models, his "congruence" restriction may seem perfectly natural: when the common term is on the right end of the existing model and on the left end of the new information, they can be superposed easily to give the combined model, otherwise they cannot. However, this is a feature of the two-dimensional nature of the **diagrams**: in three dimensions the new information could be joined just as easily to either end if the new links were perpendicular to the plane of the existing diagram.

However, Johnson-Laird would not want to suggest that models are essentially two-dimensional – indeed, he describes (Johnson-Laird, 1983. P157) mental images as views of models, which act as the source of the information from which they are generated. Instead, he regards them as very much more abstract, and in discussing them regularly mentions the properties of lists. Johnson-Laird and Steedman (1978, P77) state that "what we have in mind... is a list structure", while Johnson-Laird and Bara (1984. P29) feel it necessary to point out that "the order of the elements in a list plays no semantic role, but is relevant to the order in which elements are processed". Thus the effect of congruity arises because the properties of nodes (entities) in models are stored in a list-structured format which can only be extended by adding new information to the "loose" ends of existing lists. The result is a mechanism in which a new premise can only be incorporated into a model when the "heads" of the new nodes match the tails of the existing lists, and thus give rise to the Figural Effect.

Unfortunately, this account is far from an explanation. It relies on hypothesising that the properties of an entity are stored on a single list, in the sense of a group of items each only associated with at most one successor. Since the possibility of using nested sub-lists allows such a system to represent arbitrary information structures, it may seem that this is just a convenient way of characterising a perfectly general representation mechanism. However, a system in which items were associated with multiple successors would be equally general, and also free to allow new information from a premise to be added "perpendicular" to the old, without needing to "reverse" the model. As a result, such a system would not exhibit the figural effect.

Of course, it is tempting to argue for a "single successor" arrangement on the grounds of simplicity. However, the simplicity of a mechanism has little or no meaning without a specification of the resources with which it is implemented. Moreover, there are comparable arguments in favour of more elaborate fundamental data structures based on the (equally implementation-dependent) notion of "efficiency", which if anything have the edge on the grounds of requiring fewer sequential steps from a (known to be frighteningly slow) system. Thus Johnson-Laird's "explanation" of the figural effect stems from the imposition on the cognitive architecture of an otherwise unmotivated set of restrictions. However, this does not appear to be the case because the restrictions are so familiar from existing languages – which is, of course, precisely what Pylyshyn (1980) cautions against doing (see Section 1.4 and quotations there). If the machine on which LISP was originally developed had used a two-address machine instruction format, then there is no doubt it would have become a tree, not list, processing language, and the

true nature of Johnson-Laird's account would be apparent. By importing a familiar mechanism from computer science, Johnson-Laird is not explaining the figural effect, but he is tacitly, possibly even unknowingly, using it to set a fundamental parameter of his model.

The construction of the model also provides the mechanism for the explanation not only of the figural effect, but also the surprising (for Johnson-Laird) rise of "no valid conclusion" responses observed in Bara's "10 second" experiment. Since hurrying people produces **more** of these responses, the position adopted in Johnson-Laird and Steedman – that they are the final result of the model manipulations – cannot be maintained. Instead, Johnson-Laird argues that subjects will offer "no valid conclusion" responses when they cannot combine the models. Model building is now suggested to be a difficult and time consuming process, and when when working within a time limit of ten seconds subjects frequently fail to complete it, and hence, without any model of the whole problem from which to read a conclusion, they are forced to say that there is none. What is more, since the amount of work involved in building an initial model increases across the figures, so too should the difficulty of doing so within a fixed time, and hence so also should the tendency to offer "no conclusion" responses when the task is not complete when a response is required. In this way, Johnson-Laird and Bara explain the observed effect of figure on the tendency to give "no conclusion" responses in the "10 second" experiment. Unfortunately, this explanation too seems problematical, given that under normal circumstances people can integrate a new sentence into a discourse in very much less than a second, even if it involves bridging inferences.

Although Johnson-Laird and Bara do not mention the application of mental models beyond syllogisms, Johnson-Laird (1983) does so extensively. In addition to the specific entities and "vague" quantifiers mentioned with Steedman, the representations suitable for a range of other sentence types are described, including those relating to properties of sets. For instance, he proposes that the sentence "Scotsmen are widely scattered" can be represented in a model which represents the class of Scotsmen en masse as being identical to something which is widely scattered.

6.5. Evaluation of Johnson-Laird's Theories

Johnson-Laird's commitment to reasoning by the manipulation of situations containing or depicting distinct entities endows his theories with a powerful and natural explanation for the presence and resilience of an important feature of subjects' reasoning behaviour. As suggested in Chapter 5, there is considerable evidence to suggest that Illicit Conversion is a real error process in syllogism solving, both from modifying premises to prevent it (Ceraso and Provitera, 1971) and from attempts to warn against it (Dickstein, 1975 and Simpson and Johnson, 1966). As mentioned in Section 6.1, Wason and Johnson-Laird point out that conversion errors can be seen as arising from inappropriately assuming that the members of a set that are explicitly mentioned comprise the complete set.

Johnson-Laird's explanation for illicit conversion is essentially an expression of this idea in terms of the set representations of conceptual mental models. He argues that illicit conversion arises when bracketed items are not handled properly. Thus if the bracketed "b" in the model for "all A are B" (illustrated in Fig. 6.2(i)) is omitted when the model is constructed, the result is a

model in which the only Bs are those that were explicitly mentioned, and in which the illicitly converted premise "all B are A" is true. Notice that this means that the error is one of failing to consider all possible interpretations of the premises when they are first interpreted. If this is so, then the ineffectuality of the various attempts to warn subjects against illicit conversion is readily explained. The instructions were aimed at warning subjects against doing something illogical, whereas there is no stage at which anything illogical is done – the error is one of omission.

There must, of course, be more to a theory of syllogistic reasoning than a natural explanation of illicit conversion. However, Johnson-Laird believes that (physical) mental models, as he has previously defined them, are inadequate for the task. As a result he feels obliged to extends them, even though this is obviously an undesirable move in terms of seeking a minimal cognitive architecture. He is also aware of the necessary and intimate ties between these "conceptual models" (or indeed any other kind of data structure) and the ways in which they are manipulated, and complements their description by outlining the kinds of manipulations that he believes these models undergo. Unfortunately, closer examination reveals that neither the models nor the procedures for manipulating them are as well-defined as they at first appear.

6.5.1. The Content of the Models

Although the way they are expressed has altered, the possession of four specific features serves to distinguish *conceptual* mental models from their *physical* brethren in all versions of Johnson-Laird's theories. Each of them gives cause for some concern.

ONE The number of entities in the model is arbitrary, and signifies nothing. The central intuition behind mental models is presented as "a class is represented simply by thinking of an arbitrary number of its exemplars" (Johnson-Laird and Steedman, 1978. P77), and "although an arbitrary number of exemplars are involved, for convenience we will illustrate the various sorts of premise with a minimal number" (ibid). As a result, extra entities can be added to a model without changing the nature of the situation that it represents and indeed this is a central feature of the model-testing procedures that he proposes.

However, life is not as simple as this suggests, since closer examination reveals that the number of entities is sometimes significant. Johnson-Laird emphasises the fact that mental models can be used to describe reasoning about non-syllogistic quantifiers, such as "more than half" or "most", and to deal with precise quantification (e.g. "two"). Moreover, he even suggests that it might be significant within syllogistic reasoning. In Johnson-Laird and Bara, he discusses an alternative conclusion-reading procedure:

> Let us consider how the procedure for formulating conclusions operates when tokens of both end items are not separated by a negative barrier. **If there are more tokens** of the first end item scanned, say A, than of the second end item, C, then the procedure responds "some of the A are not C". The motivation for this procedure is simple: There are not enough C's on the same side of the barrier to be matched up with all the A's. Only when further optional C's are added in constructing an alternative model will the "no valid conclusion" be forthcoming.
>
> (Johnson-Laird and Bara, 1984. P49. Emphasis added.)

Although this suggestion only comes as part of a tentative revision of the theory, it indicates that even Johnson-Laird himself is wavering on the significance of number. There seem to be cases

where the number of entities is crucial, yet there is never any suggestion that there is anything about the model which indicates when this is so – the number of entities in any particular model is arbitrary, except when it isn't.

TWO The models in Johnson-Laird's theories contain uncertain information, represented by bracketed terms. Johnson-Laird is never fully explicit about the interpretation of complex structures involving brackets (nor indeed is he uniform in his use of them), but he clearly intends them to represent some kind of uncertainty regarding whether or not a certain individual exists (or possibly exists with particular properties). This uncertainty is central in Johnson-Laird's theories. The bracketed entities must be present to allow the models to offer a way of avoiding errors of illicit conversion, and they must be bracketed to prevent them being used to generate a range of conclusions which are not observed.

This, like the representation of an arbitrary number of entities, flies in the face of the spirit of (physical) mental models, and removes features that were necessary for their empirical support. The concept of modelling is elusive, and the closest Johnson-Laird comes to tying down what constitutes a mental model is to suggest that it is in some way isomorphic to the situation that it is representing. However, there is no obvious way in which something containing an arbitrary number of entities can be isomorphic to anything. Similarly, the key result of the experiments reported by Mani and Johnson-Laird (1982), and one of the most convincing arguments for a model-based system, is the sharp degradation in subjects' performance when the descriptions they are dealing with are forced to deal with uncertainty. The natural explanation for this is that because mental models are necessarily determinate, subjects confronted with indeterminate descriptions must abandon their preferred way of solving the problem (using models) and adopt some other approach that is able to cope. If mental models are able to represent uncertainty, this argument collapses. Of course, Johnson-Laird could attempt to start trying to distinguish types of uncertainty, but inevitably much of the force of this argument is then lost.

THREE

Mental models explicitly represent negative information, either in the form of blocked links or barriers. These are needed both to restrict the manipulation of the models and to control or prompt the reading off of negative conclusions from the models, and specifically, to ensure that negative conclusions are only generated in the presence of negative premises.

This use of explicit negative information may be tractable within the domain of syllogisms relating abstract sets. Unfortunately it has disconcerting implications for the models involved when reasoning about "real world" situations, about which the subject has a great deal of knowledge. Is the model to be seen as containing explicit negative links for everything that the subject knows is not the case? Does understanding (modelling) the sentence "all artists are Scotsmen" oblige the model builder to immediately insert negative links (or barriers) between each artist and every entity known to be of some other nationality? Not to mention noting the necessary separation between every artist and every entity that is not a human being. Recalling that, before any link can be made, the whole model must be searched to ensure that this does not result in an inconsistency, this proliferation of links clearly presents a considerable computational burden. Moreover, since there

is likely to be an enormous number of barriers between any two entities – for instance, separating those known to have the property of working in office E17 from those who don't – it is hard to see how any of Johnson-Laird's suggested manipulation schemes can be applied.

FOUR

Entity representations have some kind of internal directionality. In the 1978 theory, the description of an entity was formed from explicitly directional arrows. More recently, the directionality is no longer marked explicitly, but the layout of the model has become significant and now plays the same role. However, no matter how it is represented, the directionality is crucial for Johnson-Laird's theories since it provides the essential cue that allows conclusions to be read off in line with Johnson-Laird's own Figural Effect.

The problems with the origin of the directionality of the links as a tacit "import" of a familiar and useful construct for programming von Neumann machines was highlighted in the discussion of the Johnson-Laird and Bara (1984) account of the figural effect. Even more unfortunate for Johnson-Laird is the fact that it fits badly with the results of some of his own "cutlery" experiments (Ehrlich and Johnson-Laird, 1982. See section 3.4). While arguing against attempts to explain the Figural Effect in terms of the linguistic distinction between *given* and *new* information, he writes:

> These experiments ... were primarily concerned with quite a different matter, but one of the independent variables was the figural arrangement of the terms... Figure had no consistent effects on reading times or on memory for the premises, as reflected in the accuracy of the diagrams. **It therefore seems that figural effects occur primarily when an inference has to be made,** and in particular when a direct link has to be established between the end terms – with the middle term dropping out of the final representation of the conclusion.
>
> (Johnson-Laird, 1983. P111. Emphasis added)

Unbelievably, Johnson-Laird appears not to see the extent to which this result is at odds with his own account. Subjects in his object-layout experiments (performed with Mani and Ehrlich, and described in Chapter 3) are assumed to integrate each premise as they read it, thus maintaining a mental model of the situation as they so far understand it. The results show that figure has no effect on the ease with which subjects can perform this integration, from which Johnson-Laird concludes that "the effects of figure appear primarily when an inference has to be made". However, in the account he proposes for syllogistic reasoning, the effect of the figure of a premise is present from the moment it is understood and, moreover, it produces its effect on syllogistic error rates precisely by interfering with the **initial construction** of the mental models involved – the very process that Johnson-Laird's earlier (object-layout) results show is unaffected by figure!

In the light of these extra features, it is not obvious how the intuitions induced by rooms full of actors are relevant to interpreting the models that are so central to Johnson-Laird's theory, and his remarks concerning isomorphism do not dispel this uncertainty. Lee (1984a) offers a way of interpreting them that attempts to clarify their relationship with the situations they might be thought to represent. He stresses the need to consider the models along with the processes that revise them, and proposes a set of processes that can "generalise" each model. This involves the arbitrary replication of each entity in the model, and results in mapping each of Johnson-Laird's conceptual models into a family of models. These models are isomorphic to all and only the situations in

which the associated premise is true. As a result, they can be seen as capturing the truth conditions of the premise, which he suggests Johnson-Laird sees as crucial. Lee then goes on to suggest how the combined models can be seen as an attempt to capture the truth conditions of the conjunction of the premises, and thus of any valid conclusion.

Such a complex set of rules for enumerating the situations that any given model can encompass makes it hard not to wonder whether Johnson-Laird ought to dismiss them, as he does Venn Diagrams, as simply another "sophisticated mathematical notation" (Johnson-Laird, 1983. P93). Moreover, Lee's potential clarification of the mapping between Johnson-Laird's models and the situations they represent is not without its problems. His idea of the generalisability of each model by replicating the entities in it only makes sense if the actual number of entities is irrelevant which, it was suggested above, is by no means certain. Moreover, the procedures for model interpretation that Lee proposes do not by any means remove all the uncertainty surrounding the interpretation of Johnson-Laird's ideas, and the problems that remain are even more deep-seated than the confusions addressed by Lee.

Johnson-Laird is attempting to carry over representations expounded when discussing concrete situations (arrangements of cutlery and other familiar objects) into a problem domain concerned with more abstract concepts (syllogisms). When doing this, it is important to be clear about the degree of abstraction proposed to be present within the representation and, frustratingly, Johnson-Laird is not. As described above, he introduces the idea of using mental models for the solution of syllogistic problems via the example of ascribing roles to a room full of actors, which he initially represents as shown in Fig. 6.1. Throughout the "tableau of actors" presentation, complete lexical items (e.g. "artist", "beekeeper") are used to represent entities in models. Clearly an actor playing an artist must be distinguishable from one playing a beekeeper – the distinctions relevant within the syllogism must be preserved. But the use of complete lexical items suggests not only the ability to distinguish which of the roles in a syllogism an actor is playing, but also the ability to identify what that role is.

Johnson-Laird then suggests that it is a small step from a room full of actors to a mental model, and begins to discuss the reasoning process in terms of mental models, illustrated with diagrams, such as those shown in Fig. 6.2, where entities are represented by single letters. Clearly, there are strong practical reasons for doing this. Expositions involving "all A are B", as opposed to "all artists are beekeepers", are not only shorter but wear their generality more openly, and models can be represented more compactly. However, this change leaves one very important area of uncertainty potentially open. When considering the "standard" case of the premise "all artists are beekeepers" Johnson-Laird presents a diagram of a mental model, such as Fig. 6.2(i). Now, obviously, Fig. 6.2(i) is not a mental model. It is an arrangement of ink on a page which the reader is invited to interpret as depicting or describing a mental model. The question is *how* is it to be interpreted?

There are three distinguishable interpretations of Johnson-Laird's diagrams:-

(1) Each letter corresponds very directly to some kind of mental symbol from a small alphabet. It serves to distinguish between the possible properties relevant to any particular syllogism,

but has to be "expanded" in the context of the premises before yielding any contribution to a conclusion expressed in English. The letters in the diagram correspond almost literally to the mental model.

(2) Each letter corresponds to some kind of mental word from a very large (English-like) vocabulary. They represent the "Language of Thought" translations of the nouns in the premises. The letters in the diagrams are effectively shorthand for the internal encoding of "artist" and "beekeeper".

(3) Each linked group of symbols corresponds to some kind of model or image of an entity which is closely related to the meaning of the noun phrase used in the premises. Thus the "a = b" should be interpreted as a representation of something akin to a caricature or idealised beekeeper being somehow typically artistic.

As pointed out in Section 1.4, the content of a model, as with any data structure, is crucially tied up with the processing used in its interpretation. In this case, the details of the processing suggested can easily blur the distinction between the first two interpretations. Essentially this hinges on whether the full lexical item is used to affect the manipulation of the model, as opposed to the generation of a potential conclusion. A theory can suggest that although the model contains "letters", as in the first interpretation, these are continually being expanded or interpreted according to the premises. In this case the theory is effectively supporting the second interpretation – the model may as well just contain the words. Similarly, if a theory that claims that its models contain full lexical items, but suggests that they are processed in such a way that the meanings of the words are not accessed, it could just as well have only letters in its models. Essentially the "symbol" interpretation suggests that model manipulation proceeds independently of which words are employed in the premises, whereas the "word" interpretation suggests that it is affected.[10]

These different interpretations of the diagrams lead to differences in the mental models they suggest, and thus to differences in the mental processing that we are to believe are happening. These differences are, of course, highlighted in the case where the actual lexical material influences subjects' reasoning. Admittedly no experiment has yet been run to present subjects with such semantically biased syllogisms to **solve** (as opposed to verify – Wilkins, 1928). However, there is good reason to believe that such influences will exist. Johnson-Laird himself has demonstrated their presence in other reasoning tasks (e.g. using the Wason "four card" problem), and has suggested that it is a significant flaw in the "standard", mental logic, explanation of reasoning that it cannot explain them (e.g. Johnson-Laird, 1983. P29 - 34 and numerous references there). Moreover, experience of discussing syllogisms with untrained people (e.g. subjects) shows that it is easily possible to use semantically biased examples **with no further explanation** in order to point out errors. Subjects can readily tell, from a semantically biased example alone, that there is something wrong with a conclusion. Thus we have every reason to expect that syllogistic reasoning will be affected by the content of the problems.

---

[10] Steedman (personal communication) has suggested that earlier presentations of the 2nd and 3rd interpretations were indistinguishable. This will be dealt with below.

Given that an effect of material on performance can be confidently expected, we can reasonably examine how each of the three interpretations of mental models can explain it. First, consider a simple case such as:

all women are hill walkers.
some hill walkers are mountaineers.

As explained above, Johnson-Laird suggests that subjects will interpret this syllogism by building a mental model, which he would depict as shown in Fig. 6.8(i).[11] On the basis of this model, subjects are tempted to draw the conclusion that "some women are mountaineers". However, a more diligent subject will proceed to test their mental model, and discover that it can be rearranged to a form printed as shown in Fig. 6.8(ii). This falsifies the previously supported conclusion, and such subjects will therefore say that there is no valid conclusion, which is the correct response. Syllogism experiments do indeed show that subjects are divided between these two responses (and virtually no other) thus confirming Johnson-Laird's predictions.

Now consider the implications of changing the lexical material, replacing "mountaineers" by "men". For the simple syllogism, this leaves the theory with the embarrassing possibility of predicting that a large proportion of intelligent adult subjects will suggest the conclusion that "some women are men". Common sense and the evidence outlined above suggest that this will not be a common response. Thus we expect that a theory of reasoning will indicate how the influence of the material comes about – why we will not see the obviously nonsensical conclusion. Considering this problem highlights the differences between the three interpretations of Johnson-Laird's diagrams.

In the first approach mentioned above, the letters in the diagrams denote symbols which merely select one of the properties involved in the syllogism.[12] Under this interpretation, the mental model built in response to the modified (men) premises is **identical** to that built for the original (mountaineers) premises. It must be, since the premises themselves are semantically unobjectionable, so there are no grounds for postulating variable behaviour at the premise encoding

Fig. 6.8: Modelling a Semantically Loaded Syllogism

[11] This is in the post-1984 notation, though the argument is the same using the earlier exposition of the theory.

[12] As pointed out, this is also equivalent to a lexical system where the meanings of the "words" are not referenced.

stage. Similarly the same manipulations either do or do not take place in both cases, because this model interpretation is characterised by the suggestion that the model is composed of tokens that are independent of the lexical material of the syllogism. This means that the unwanted conclusion can still be read off the model and therefore, in order to explain any semantic effects, it is necessary to suppose that the solutions generated by the mental models are somehow "vetted" by some other, content-sensitive, mechanism which would recognise implausible suggested conclusions and reject them.[13] Apart from being conspicuously unsupported by either intuition or anything said by Johnson-Laird, this task presents a formidable processing requirement for anything but a syllogistically trivial situation.

Consider now the second possible interpretation suggested above: that the diagrams denote mental models that are to be thought of as a collection of links between linguistic items of some kind. What distinguishes it from the first approach is that the semantic import of these linguistic items influences the process of model manipulation. Thus both the models built and the conclusions that can be read off them can be affected by the lexical content of the premises. In the case of the examples under consideration, the models of Fig. 6.8(i) and 6.8(iii) would never be built, because the impossibility of describing any entity as both "man" and "woman" would prevent the links encoding the last premise being made. This means that there is no need to postulate any kind of conclusion-vetting mechanism. However, doing so brings with it the computational burden of checking the compatibility of a number of pairs of descriptive terms that rises with the number of entities every time the model is amended.

Both these interpretations of the diagrams call for a mechanism for the recognition of contradictory descriptions, such as "a male princess", though they differ in when it is to be applied. However, having called for such a mechanism, it is still necessary to consider its operation. The most common formal technique for the manipulation of meanings of words would involve solutions couched in terms of some kind of *meaning postulate* (Carnap, 1947): I.e. explicit stipulations of the implications and incompatibilities of words, such as "man(X) => ~ woman(X)". However, a simple collection of such postulates is clearly inadequate, since everybody recognises that there is an (effectively) infinite list of things that, for instance, princesses are not: dustmen, wasps, kingdoms, natural numbers...

One can attempt to avoid this infinity by recognising that, in a very strong sense, "some princesses are men" is anomalous **because** princesses are women, which suggests that some kind of type hierarchy might be appropriate – that is, some method of representing the interrelations

[13] The problem can be made still more acute by simply extending the argument with the additional premise that "all of the athletes are women". Johnson-Laird's own experiments indicate that almost all subjects will successfully derive an A conclusion to a **IAA**, so the additional premise should not significantly affect their performance. Thus subjects can be expected to approach the last premise in much the same state as they would in the simple case, and many of them can be expected to offer the conclusion that "some of the athletes are mountaineers". In the semantically loaded instance, however, the addition of a conclusion vetting mechanism cannot save the theory from having to make the counter-intuitive prediction that subjects' behaviour will be unaffected. There is nothing anomalous about the conclusion, and therefore nothing that any kind of conclusion-vetting mechanism can detect. This leaves the embarrassing prediction that they will notice no anomaly, and happily infer something – "some of the athletes are men" – that flatly contradicts one of the premises. In fact, there is nothing anomalous about any sentence involved in the deduction, and the problem could be detected by nothing short of verifying every sentence that could possibly be read off the model – not just those relating end terms.

between category words. This approach would also be needed to explain a different, almost opposite, influence of the lexical material: the utilisation of unstated premises. The following deduction will surely cause nobody any problems:

Dizzy is a dalmatian.
All dogs are mortal.
therefore
Dizzy is mortal.

However, it is beyond the models suggested by the interpretations we have considered so far, because it relies on the recognition of the implied premise "all dalmatians are dogs", which in turn requires that a type hierarchy be searched, this time for all possible class inclusions. What is worse, once this is done the new descriptions must in turn be checked, both for further class inclusions and incompatibilities. Worst of all is that at each stage the process must take account of arbitrary amounts of information, possibly involving multiple type hierarchies simultaneously: Pooh Bear is undeniably a bear, and bears are indeed all mortal, but, unlike Dizzy, Winnie the Pooh will never die!

It seems that the first two possible interpretations require a mechanism that can detect the incompatibility of lexical items and access unstated premises. Trying to do this using any kind of type hierarchy requires a mechanism able to simultaneously handle multiple hierarchies in combination with arbitrary knowledge about the world. Such a reasoning engine is a very unsatisfactory thing to have in a theory based on mental models. Johnson-Laird makes no mention of any mechanism remotely like it, and indeed it exemplifies the very approach that he intended mental models to replace.

So far, it has been argued that the first two interpretations of Johnson-Laird's mental models flounder when trying to allow for the effect of "leading" material on subjects' performance, and force us to postulate mechanisms that are dissonant with the ideas of a theory based on mental models. The discussion in Chapter 3 of the interaction of language and models suggests that there may be a way of detecting the incompatibility of descriptions that is more in line with the aims of mental models. There it was suggested that inference is avoided by building an appropriate model, which will automatically fit with any information implicit in what has been said, and, if necessary, examining this model to **discover** any consequences that this information might have.

Using this approach, it is possible to extend the first two interpretations more in line with the ideas of mental models by proposing that subjects detect inconsistent descriptions because they are unable to build a representation of the object described. One could suggest that before any conclusion was accepted (under the first interpretation of the diagrams) or any link constructed (under the second), the compatibility of the terms involved would be checked by attempting to model or imagine an entity to which both could be applied. This obviously requires that it is impossible to create something that is both a man and a woman, but this is assured, since the ability to create such an object would mean it was possible to conceive of something in the world (i.e. something that the model modelled) to which both the terms applied, and this is by definition impossible for contradictory terms. Similarly, the combined model, once constructed, would be used to identify implicit premises.

In this way, the first two interpretations of Johnson-Laird's diagrams can be seen as **compatible** with the ideas of *Mental Models*. However, there is undeniably something perverse in the suggestion that the operations of the (conceptual) mental models are validated by some other process that involves creating other (physical) mental models! Not only does this sound like a definite distortion of Johnson-Laird's theories, but what is more the "secondary" mental models have all the properties of those of the primary mental models under the third possible interpretation: caricatures or idealisations of the things they represent. This suggests that it is the third of the possible interpretations that is most in keeping with the spirit of his enterprise and the goal of explaining syllogistic reasoning behaviour without recourse to mental logics.

It should by now be obvious how the third, preferred, interpretation of the mental model diagrams would predict the effect of the content of the syllogisms. The mental model of the first premise ("all women are hill-walkers") would be akin to some kind of picture of a row of women, all dressed for hill-walking, with perhaps some other (male) hill walkers in the background. However, when the subject attempts to integrate the second premise the effect would differ according to the content of the premise. In response to "some hill walkers are mountaineers" the subject might imagine some of the hill walkers (including some who were women) clutching ice axes and wearing crampons and ropes, or possessing some other features that the subject associates with mountaineers. This would correspond to the first diagram proposed by Johnson-Laird, and would thus lead to the prediction of the erroneous responses of "some women are mountaineers".

On the other hand, to treat the alternative second premise "some hill walkers are men" in the same way would require the subject to give some of the women features that they associated with men, such as beards or deep voices. But to do so would be impossible, because these new properties would conflict with the properties they have already been given, and thus the subject would baulk at building the model they built for mountaineers. To make the premise true, the subject would have to create some **other** hill walkers who were men and would thus go straight to a model represented by the final diagram (Fig. 6.8(ii)) that Johnson-Laird proposes.

We can see here the core of the reply to Steedman's suggestion that the 2nd and 3rd alternatives are indistinguishable. He argued that both can be seen as representing entities that have certain properties, with the linguistic items tagging a node simply acting as pointers to the concepts that the corresponding object is modelled as having. However, as the discussion of physical models should have made clear, there is an essential difference relating to properties not explicitly mentioned. If a "true model" approach is adopted, the object modelled will implicitly acquire certain additional properties in the process of being given the properties explicitly mentioned. On the basis of being told that "Pluto is a pet that barks", a mental model of Pluto will incidentally be a model of a dog. In contrast, in a linguistic system, he will remain only a canine pet until the probability of his doghood is explicitly deduced from the knowledge that pet wolves are rare.

It has been argued, therefore, that the third interpretation is the one that is most in line with the ideas and intuitions of mental models. It is also the only one that will allow for a natural explanation of the effects of material and background knowledge on the reasoning behaviour of

subjects. It achieves this because the model of any entity represents not only the properties that the premise identified as salient, but a range of other features drawn from the subject's background knowledge of the kind of entity involved. It is the construction of that homogeneous representation of mentioned and implicit properties that both allows and controls the influence of the material being reasoned about. However, there is a sense in which it is deeply incompatible with the properties of the conceptual models that Johnson-Laird suggests.

It is only sensible to speak of directionality and explicit negation when one is referring to inter-relations between lexical items, or some other identifier of isolated and specific properties. Therefore these features of conceptual models can only be found in models such as those described under the first two interpretations, which as has been shown require the support of some kind of logic engine in order to function. Moreover, such a reasoning mechanism could only be applied to an everyday situation if it were preceded by an explicit recognition and lexicalisation or abstraction of the salient features of the situation. This is equivalent to requiring the explicit recognition that transitivity is appropriate to a spatial arrangement. Using Johnson-Laird's own example, if syllogistic reasoning using mental models plays any part in determining how "Person A" decides to ask the group of people how to get to the University, it can do so only after "Person A" has first explicitly formulated and lexicalised the two premises. Thus it seems that making the decision was the result of a process that involved a complex inference engine and the combination of two explicitly represented and believed-to-be-true sentences. But such a view is surely the paradigmatic case of "logic in the head" that Johnson-Laird was wanting to avoid.

6.5.2. Processing the Models

Johnson-Laird is aware of the impossibility of saying anything worthwhile about a mental model, or any other data structure, in isolation from the procedures that manipulate it. This leads him to outline the kind of procedures he believes are involved with the initial building and interrogation of a mental model, as well its manipulation.

Since each individual premise is mapped to a unique mental model, the procedures for model construction are straightforward. As a result, they are not dealt with beyond suggesting that, since syllogistic premises are sentences of everyday language, their interpretation into mental models can be assumed to be done by the subject's linguistic apparatus.

> The process of forming an integrated mental model of premises is nothing more than the proper comprehension of discourse: It is required in order to grasp the full impact of what a speaker has to say. The ability to carry it out should be common to all native speakers of a language, and, since it and its complementary skill of putting models into words suffice for competency with syllogisms requiring only one model, it is hardly surprising that the subjects were almost universally able to cope with these syllogisms.
>
> (Johnson-Laird, 1983. P119)

Johnson-Laird is a little more forthcoming about the production of conclusions. Because he sees the mental model as playing a central, active, role in the reasoning process, he suggests that the generation of a conclusion must involve reading it from the model (or set of models, for many-model problems). Johnson-Laird (1983, P34) points out that there are an infinite number of (mostly

trivial) things that can be said about any situation,[14] and he describes (ibid, Chapter 3) a system which (for the most part) generates only the kinds of statement that are actually made. In the context of syllogism solving, this takes the form of a mechanism that always reads off the strongest simple relation that holds between the end terms in all the models being considered. Correct reasoning is assured by tailoring the models to ensure that only one or two conclusions are ever considered.

The mechanism Johnson-Laird proposes is, of course, necessarily imperfect. Incorrectly saying that no conclusion can be drawn is an common error, which is infrequently but regularly observed even with first figure premises. However, Johnson-Laird can only explain this failure to read a conclusion as some kind of performance error. The most common response to **4AA** syllogisms (an **A** conclusion) is similarly inexplicable, since it cannot be read from any of the models that Johnson-Laird suggests are built! But the least pleasing feature of the system is the fundamental difficulty of extending it to other domains. Obviously the mental model will represent what can be said (i.e. what is believed to be true) in any situation. However, the core of Johnson-Laird's ideas on conclusion generation is that the model to a large extent constrains the range of candidate conclusions: much of the notation of conceptual models is geared towards indicating what should be said or how. While this specificity may be feasible for syllogistically trivial examples, there is no reason to suppose that it can always be achieved.

To see this, consider the fact that in their ordinary lives people freely produce an amazing range of linguistic behaviour. Even in a trivially simple situation such as blocks-world (Winograd, 1972) there is an enormous range of things that a person could say (infinitely many, in the case of a logician) even if they chose to limit themselves to true statements, which is by no means a universally observed policy. Restricting this choice – deciding what to say and how to say it – is a skill. Of course a mental (discourse) model has a central role to play in this process, as an encapsulation of the cognitive system's information concerning what is believed or supposed to be the situation under consideration. But conversing is a complex process, making subtle use of information quite unrelated to the subject matter being discussed: the status and mood of the person being addressed and the social situation of the conversation, what the addressee already knows, the objective of the utterance (what Austin (1962) termed its *perlocutionary force*), what has already been mentioned etc.

Given the complexity of the task and the diversity of the relevant information, there is no reason why the mental model should be expected to bear the burden of constraining the choice of what should be said beyond, of course, indicating what could be said. Thus Phil's social situation (more specifically, in conjunction with his social intentions) undeniably affects whether he describes his home as having a lavatory, a loo or a toilet, and the progress of the conversation determines whether it could be referred to as "it". However, these effects are surely achieved by influencing the lexicalisation strategy Phil adopts, and not by altering his well established and stable mental model of (i.e. his ideas about) where he lives. Indeed, the standard arguments against

[14] at least in the sense of being sentences which are true of that situation.

behaviourism based on linguistic behaviour rely on just this idea: while undeniably guided by it, a person's linguistic behaviour is not constrained by the situation they are discussing.

On the subject of how the model modification process is organised, Johnson-Laird and Bara initially suggest a "straw man" approach. An adequate model modification procedure could involve repeatedly making random modifications to the model, and then checking whether the tentative conclusion had been falsified while still satisfying the premises. However, they then point out the computational inefficiency of such a process, and describe (pp 38 - 40) a set of rules that they have used in a computer implementation of the theory. These rules concentrate on the effects of the "barriers", are complex and appear ad hoc. In particular, the rules must distinguish properties invoked by syllogistic end terms from those invoked by the middle term, and the permeability of the barrier to any individual depends on other entities with properties in common with that individual. Thus an entity which "optionally" has a property ascribed by an end term can freely pass a barrier, while those that definitely have it cannot, although they can do so when two barriers are involved, even when this involves creating a model which no longer supports both the original premises (e.g. the **1EO** syllogism). In contrast, entities that optionally have the property ascribed by the syllogistic middle term can never simply move past a barrier, although they can "swap" with non-optional middle terms, and this mechanism can allow an entity definitely described by an end term to pass a barrier when it otherwise could not.

Having presented these rules, however, Johnson-Laird and Bara go on to say:

> Obviously, we do not wish to defend the psychological reality of these procedures. We do not know what procedures people use to manipulate models; we suspect that they are neither as haphazard as the purely random method described above nor as systematic as our five procedures. The crucial point is the number of models that have to be constructed by them in order to draw the correct conclusion.
>
> (Johnson-Laird and Bara, P40)

This quotation suggests that there is a sense in which (or level at which) Johnson-Laird does not see himself as putting forward a clear statement about how syllogistic reasoning problems are solved, but as trying to outline the **kind** of approaches are used. In this case, the details he presents should be seen as merely one (comparatively) arbitrary way of filling in the outline, and the theories he presents should be taken as representative members of the **family** of approaches that he is supporting. Such an idea would certainly be in keeping with the way that Johnson-Laird's theories have been evolving over the years. Is Johnson-Laird's latest theory to be seen not as his best approximation to how syllogisms are solved, but as his clearest pointer to the kind of phenomena that he believes are involved?

Unfortunately this interpretation is at odds with the features of the theories that Johnson-Laird presents. The difficulty of a syllogism is measured by the number of models that it requires, and the range of erroneous conclusions offered depends on what can be read from them. Both of these are closely dependent on the manipulations that can be employed in moving between models. Consider the **1EE** syllogism. According to Johnson-Laird, this is a two-model problem with the models depicted in Fig. 6.7 and reproduced, for convenience, in Fig. 6.9. This allows him to predict, correctly, that subjects will either recognise that there is no valid conclusion, or they will offer an **E** conclusion. No other conclusions are predicted or observed. However, this prediction

Fig. 6.9: 1EE syllogism: Johnson-Laird's models.
(This figure is identical to Fig. 6.7)

depends crucially on the details of the manipulations that Johnson-Laird proposes. The moving of a pair of end terms past a barrier constitutes a single transformation: if it did not, and the terms could move one at a time, then the problem would be a three-model problem, and **O** conclusions could be predicted.

Similar problems arise in other places. Johnson-Laird's three-model problems all share a common history: an original model is modified to create a second model by moving a single optional end term past a barrier, and then a third model is created from this by duplicating this optional end term. Thus this replication of an entity increases the number of models. The difference between the second and third models is that the second will support invalid "O" conclusions that the third will defeat. Since between the 12 three-model problems these invalid conclusions are only given twice in Bara's data, the empirical evidence for the existence of the second model is scanty, to say the least. But if the replication of the entities did not increase the model count, the three-model problems would only be two-model problems. Similarly the duplication of an entity is all that distinguishes the two models in the **3OA** syllogism. Were it not significant, it would be a one-model problem, albeit an extraordinarily difficult one, and once again the empirical support for the smaller model is minimal – only 1 subject, which Johnson-Laird would have ignored if it were not a predicted response.

In contrast, when a single optional end term moves past two barriers, it can be replicated in the same operation. This means that the **OE** syllogisms are all two-model problems, whereas if the duplication there were recognised as a separate step many would have to be classed as three-model problems, albeit ones that most subjects solved correctly. There is no obvious reason why optional end items moving past barriers cannot be replicated while their non-optional brethren moving past two barriers can.

The above cases strongly emphasise that there is no clear rationale behind the apparent unitary operations that subjects can perform. However, the classification of syllogism difficulty in terms of the number of models required, which Johnson-Laird's theory stresses so much, depends completely upon these apparently ad hoc features. Thus it seems that Johnson-Laird cannot be proposing a

family of theories. Any other member of the family will differ in its classification of syllogistic problems, and this will thus completely undermine the empirical foundations supporting his account – namely the prediction of the difficulty of each item. The details of the theory are important, and Johnson-Laird's colours are nailed firmly to their mast.

Nonetheless, Johnson-Laird and Bara's denunciation of the details of the procedures is a significant statement, since it also casts doubt on many of the central features of the theory, and in particular on the need for conceptual models themselves. The approach to model manipulation they offer as a "straw man" (random modifications with consistency checks against the premises) did not make any use of the presence of the barriers within the models – properties could be moved freely, and inappropriate changes were rejected when they were found to falsify a premise. Thus the only necessity for the barriers is for the cueing of the generation of negative conclusions. Similarly, if any property might be considered for deletion, there is no need to distinguish positive relationships from uncertainty: an optional relation is one that can be deleted without violating the premises. The only processes which need the explicit representation of uncertainty are once again those involved in the generation of conclusions, which must not utilise uncertain information.

Admittedly these are not the arguments that Johnson-Laird puts forward to support the use of sophisticated conceptual models over making random modifications then verifying the premises. He argues that the latter is implausible because of its inefficiency,[15] though since subjects seem to think that solving syllogisms is difficult, this may be no bad thing. More importantly the process may, in fact, never finish,[16] and if terminated externally, might well not have considered all possible model modifications (i.e. possible counter examples). Given subjects' widespread tendency to accept invalid conclusions, Johnson-Laird might regard this, too, as a feature, were it not for the fact that it makes subjects ability to reason correctly (their "rationality") a matter of chance. However, all these problems arise from the fact that the modifications postulated are **random**. Even supposing it is meaningful to speak of people behaving "randomly", there is no reason why this should be the case. If this problematical style of proceeding is replaced by any kind of methodical approach, these arguments collapse.

The extensions that distinguish conceptual models from physical models are complex and hardly intuitive, and parsimonious theorising demands that we challenge them vigorously. By countenancing a model manipulation procedure akin to their straw man, Johnson-Laird and Bara reveal the true reason that physical models are deemed inadequate for syllogistic reasoning. It does not lie in the **logical** processes of comprehending and considering the situations described by the premises, where their complexity might seem justified as a means to provide a mechanism for performing inference without logic. Instead, it lies in the generation of suitable conclusions. The observations they explain are closely akin to those described by the much simpler, even simplistic, Atmosphere Effect. Moreover, the way they explain them is conspicuously inextensible. Seen in

[15] Since several easy steps are often preferable to a few hard ones, discussing what is "efficient" in the brain presumes some way of estimating how difficult any particular task might be. Certainly the amount of computation required by a digital computer is no measure – compare the relative difficulties of hearing a sentence and finding a square root.

[16] this would tend to make it inefficient :-)

this light, they appear both cumbersome and ad hoc, and Occam's Razor becomes a Sword of Damocles.

6.5.3. The Origins of the Mechanism

According to the most obvious way of reading his theorising, Johnson-Laird is putting forward a notation set up to deal with arbitrary numbers of members of sets, manipulation mechanisms for modifying them in line with syllogistic premises and others for generating syllogistic-type conclusions. This represents a very syllogism-specific mechanism, and he gives no indication of how it could have arisen. Given that he thinks it is part of everyday reasoning, he would presumably adopt the same approach as is tacitly adopted by other theorists, suggesting that it arises by some combination of genetic engineering and everyday experience – that it can be thought of as a "natural" ability like, or even akin to, language.

However, there is a problem with this suggestion: subjects' performance changes noticeably as they solve syllogisms. If the mechanisms underlying reasoning are the result of genetics and everyday experience, then an experiment presenting only 64 abstract problems with no feedback would constitute only a negligible proportion of the tasks to which they had been applied. Thus, particularly in the absence of feedback, it would be most unlikely to teach the reasoner anything or to bring about a change in behaviour. However, this is not the case: solving syllogistic problems does change subjects' solving ability, even within the very limited experience of a single experiment. Some evidence for this comes from the differences between subjects' performances on the two "sittings" in the experiment that Johnson-Laird and Steedman (1978) report. They point out (P75) that "there was a distinct improvement in performance from the first test to the second test", with 19 out of 20 subjects improving, and 44 of the 64 syllogisms being handled better. More evidence comes from Inder's experiment, which suggests that a subject's behaviour changes not just over a series of several sessions (see Section 6.2.2.9), but even while solving a single set of syllogisms (see Section 7.3). However, Johnson-Laird's theories say nothing about how or why these changes should occur.

Johnson-Laird also ignores another factor, closely related to subjects' changing behaviour: the variability of results between subject populations. There is a qualitative difference even between the results of Johnson-Laird's own experiments, on which he himself comments (Johnson-Laird and Bara, 1984. P23). The subjects in Bara's experiment offer many of what Johnson-Laird terms "Gricean" conclusions – conclusions that are not valid inferences from the premises, but are validated by the attendant Gricean implicatures. In contrast, the results reported in Johnson-Laird and Steedman (1978) give no indication of any inclination to produce such conclusions. Furthermore, such differences have been evident from the very beginning of work on syllogistic reasoning. When Chapman and Chapman (1959) first tried to replicate the work of Woodworth and Sells (1935), there were obvious features (the presence of **E** conclusions for **EE**, **OE** and **EO** syllogisms) of the observed distributions that could not be ascribed to differences in experimental method.

These variations appear between the results of groups of subjects, so cannot be ascribed to performance errors or differences in individual ability. Furthermore, there is no reason to suggest

that these widespread variations can all be attributed to unreported features of experimental method. This implies that they reflect real differences in the way the different populations solve syllogistic problems. These must have arisen as a result of differences in their cultures leading to different experiences in the relevant areas.[17] However, as Johnson-Laird presents them, his theories postulate a tightly integrated and highly interdependent syllogism solving system, with very little scope for modification compatible with the paramount objective of explaining rationality. He offers no indication of the "degrees of freedom" that could explain the observed performance variations, and there are no features which appear likely to be in any way culturally determined. The theory can account for differing populations only by ad hoc modifications – i.e. not at all.

This inability to deal with temporal changes and the variations between different groups of subjects reflects a fundamental feature of Johnson-Laird's theorising. The essence of his approach is a system in which subjects pass through a number of stages on their way to a solution to the problem. This mechanism will always give a correct answer if run to completion, since otherwise it could not support correct reasoning and "rationality". This means that the only way Johnson-Laird can explain both differences between subjects and the abundance of imperfect reasoning is by suggesting that they sometimes abandon the solution procedure at intermediate stages. Since experimental subjects make reasoning errors on approximately 50% of the syllogistic items, these errors are clearly an important determiner of actual behaviour and thus worthy of close attention. In particular, since the different classes of error identified in Section 4.5 have different properties, in particular when the theory is extended to non-syllogistic problems, it is significant to decide to **which** category of error they belong.

Johnson-Laird himself suggests that a satisfactory theory of syllogistic reasoning must be able to explain how it is that people can be rational. He believes that the experiments probe the fundamental mechanisms of thinking, so he cannot ascribe the observed reasoning errors to competence limitations. To do so would be to suggest that people are fundamentally incapable of carrying out those deductions correctly – that logically sound thinking is beyond them – and this is clearly not a move he would wish to make. This belief also raises other problems, since suggesting that subjects' procedures of everyday thought are beset by ability errors implies that their conclusions are de facto always unsound. Further, he says nothing at all that could suggest he is thinking in terms of ability errors, and his statistical approach to errors confirms that he is thinking in terms of some combination of capacity and performance errors.

Johnson-Laird and Bara identify two features that make some syllogisms harder than others – the number of models required to solve it correctly (also highlighted by Johnson-Laird and Steedman), and the figure. Each of these factors is thought to increase the amount of work that correct solution requires, and this is alleged to give rise to the highly significant trends that show up in the patterns of correct conclusions in Bara's data. This support, of course, rests on the assumption that the more processing the subject does, the more likely they are to be affected by a

[17] Or possibly their genetic makeup, though subtle differences in syllogistic reasoning do not seem to justify opening that can of worms.

performance (i.e. execution) error and thus make a mistake or give up.

Johnson-Laird and Steedman directly attributed the effect of the number of models to execution errors. The more models that have to be processed, the more chance there is that the subject will give up part way. Specifically, they will either prematurely abandon the conclusion-testing loop and accept a conclusion that could be defeated, or they will fail to "scan the model in both directions", and miss a conclusion that runs counter to the figural bias of the model. Similarly, for Johnson-Laird and Bara the effect of figure arises because the models for the premises can only be combined when their terms are suitably ordered (the mechanism that explains the figural effect). For each figure the premise models have to be manipulated differently in order to reach a suitable configuration, and since this manipulation introduces extra processing, it is also predicted to introduce extra (opportunity for) error. Specifically,

> The complexity of the operations required to integrate the premises increases over the four figures. Hence, the proportion of correct valid conclusions should decline over the four figures with a correlated increase in the number of 'no valid conclusion' responses and errors.
>
> (Johnson-Laird and Bara, 1984. P34)

Hence they predict that the figures will be ordered in terms of difficulty: $1 < 2 < 3 < 4$, with the errors manifesting themselves in subjects incorrectly saying that there is no definite relation between the end terms.

These predicted trends in difficulty are confronted by a particularly conspicuous problem even within Bara's own data. The **4AA** syllogism is a 1-model problem with a symmetrical (**I**) conclusion that cannot be missed by "reading the model in the wrong direction" because of the figural effect. Thus it should be one of the easiest syllogisms, certainly within the 4th figure, and yet not one of Bara's subjects offered the right conclusion, and Inder and Steedman both show poor accuracies on this item too. Admittedly the fourth figure is predicted to be the most difficult, but not to the extent of eliminating correct conclusions. Moreover, the fourth figure is predicted to be troublesome because subjects will have difficulty building a combined model, which will cause them to incorrectly say that there is no valid conclusion, which is not the kind of error that Bara's subjects make (they offer **A** conclusions). Johnson-Laird and Bara thus have no way of explaining their abysmal performance except as a result of a grossly uneven distribution of performance errors, which, as argued above, indicates a significant factor is being overlooked. A much more satisfactory way of accounting for the performance is to suggest that there is something particularly difficult about the **4AA** syllogism, but no such suggestion is forthcoming.

Even if we set aside the **4AA** syllogism as an isolated shortcoming of the theory, there are still a number of results, even from Johnson-Laird's own syllogistic reasoning experiments, that mean that these two causes of difficulty deserve further attention.

The trend of increasing difficulty with number of models is general: it is clearly supported both in the results of Inder and Steedman. It is also to be expected. As pointed in Section 4.2.2, the search space of syllogistic problems has two "degrees of freedom" – the ambiguity inherent in each of the premises and the indeterminacy in the way that they are combined. As Lee (1984a) points out, Johnson-Laird's conceptual models directly represent the former – each premise has a unique representation. Therefore the number of models that a syllogism requires can be seen to be

a direct measure of the ambiguity of combination. However, Ceraso and Provitera (1971) used syllogisms with disambiguated premises, to establish that the ambiguity of combination is indeed a source of errors. Thus the effect of number of models that Johnson-Laird observes could have been predicted from the earlier results. Moreover, the correlation of the number of models that a syllogism requires with error rates may tell us nothing about the mechanisms employed in its solution. Because it reflects the amount of ambiguity of combination, it is a measure of one dimension of the search space of the problem, which in turn can, at least prima facie, be expected to be a factor contributing to its difficulty. Problems requiring many models provoke more errors because they are harder.

While recognising that the number of models represents a measure of problem difficulty would by itself lead one to expect it to correlate with error rates, Johnson-Laird attempts to make a more powerful case. He suggests that

> The search for alternative models is likely to place a considerable load on working memory, the greater the number of models to be considered, the harder the task should be.
>
> (Johnson-Laird and Bara, 1984. P40)

and that

> Subjects do indeed attempt to assess alternative models of the premises, but often the task **exceeds the capacity of their working memories.**
>
> (Johnson-Laird, 1983. P105)

This is suggesting that the correlation of the numbers of errors and models is caused by "working memory" limitations – that it is a *capacity* error. Assuming that people can only work with a limited amount of information at any time, he is suggesting that three model problems, of which Bara's subjects get only 3% right, are coming close to this limit.

The gross trends in Bara's data support the idea that 3-model problems are very difficult because they induce capacity errors as the limit of working memory is approached. The total numbers of subjects solving 1-, 2- and 3-model problems fall steadily. However, capacity errors can be expected to give specific and reliable distribution, and the Steedman data certainly do not fit at all well. Most 3-model problems are handled correctly by between 13 and 17 of his 20 subjects, which suggests that they found them easier than many 2-model problems, 5 of which were correctly handled by fewer than 9 subjects. Similarly, Inder's subjects handled several 2-model problems less well than some 3-model problems, while even Bara's subjects find some 2-model problems more problematical than some 3-model problems, and others easier than some 1-model problems. It may well be that 3-model problems are more difficult, on average, than 2-model problems, but all things being equal, capacity errors will produce a uniform distribution of difficulty. These anomalies undermine Johnson-Laird's attempt to ascribe the effect of figure to a capacity limitation. They can only be described as execution errors with an curious and unexplained distribution – once again suggesting that some other factor is involved.

Of course, the number of models associated with any particular syllogism form is a result of the range of model manipulation operations that Johnson-Laird has chosen. As a result it is tempting to suggest that, in the absence of arguments for why these precise procedures are correct, the fit to the data could be improved by selecting some other set. Indeed, given that there are an

infinite number of possible procedures from which an undetermined number must be chosen, it seems quite likely that it is possible to find a combination that fits the data better than those that Johnson-Laird proposes. However, a post hoc search for such a combination simply becomes an exercise in parameter setting, the worth of which is as unclear as their relevance beyond the particular laboratory task on which they were assessed.

The situation regarding the effect of figure on error rates is also confused. Johnson-Laird suggests that

> Since the complexity of the operations required to form a mental model increases over the four figures, there should be a corresponding increase in the difficulty of drawing a valid conclusion, and in the latency of responses.
>
> (Johnson-Laird, 1983. P110)

while with Bara he literally emphasises that the effect is on valid conclusions. However, if it has this effect by preventing the formation of a model which is essential to the drawing of a conclusion, it is not obvious why it should not affect **any** conclusion, valid or otherwise. Thus it should predict an increasing tendency to offer "no valid conclusion" responses, which in turn would lead to an increase in the total number offered. The position with the fourth figure is complicated because it presents significantly more opportunities to draw a valid conclusion – in the case of subjects who are (hopefully) influenced by the logic of the situation, this should lead to a reduction in the number of "no conclusion" responses offered. Nevertheless, the trend at least ought to be observable in the first three figures, a prediction clearly confirmed by all the groups of subjects (see Table 6.6). Thus figure clearly does affect the tendency to draw a conclusion.

However, other results are more troublesome. Because he considers only valid conclusions, Johnson-Laird is led to a more subtle prediction that should apply equally in the fourth figure: subjects accuracy on syllogisms which have valid conclusions will be degraded. For each group of subjects, Table (6.6(a)) shows the percentage of correct responses to items that allow valid conclusions to be drawn (the factor which Johnson-Laird uses to support the effect of figure). Bara's results clearly show the predicted trend, and those of Steedman1 are at least compatible with it, but the results of Inder and Steedman2 confound it – both groups of subjects do significantly better in the fourth figure than the third. Furthermore, since the predicted effect of figure is to bias subjects to say that no conclusion can be deduced, this degradation in accuracy should be accompanied by an increase in the percentage of (wrong) "no valid conclusion" responses to those same syllogism. Table 6.6(b) shows that no group of subjects clearly shows this trend, and once again Inder and Steedman2 are contrary to it.

The increasing frequency of "no valid conclusion" responses across the first three figures, in all the sets of data, supports Johnson-Laird's suggestion that the figure of the premises influences subjects' willingness or ability to offer a conclusion. So too does the observation that, with the exception of Bara's 3rd figure, the sum of the corresponding percentages for valid conclusions and "no valid conclusion" responses – and thus the percentage of incorrect conclusions offered – is more or less constant for each set of data. This suggests that the main influence of figure is to switch subjects between giving the right conclusion and saying that there is none. However, this influence of figure is by no means simply a universal tendency – the result of some fundamental

(a) Percentage of Responses Drawing Valid Conclusions.

Experiment	Figure: 1	2	3	4
Bara	48%	45%	33%	22%
Steedman1	60%	53%	53%	49%
Steedman2	70%	60%	60%	69%
Inder	53%	53%	39%	49%

(b) Percentage of Incorrect "no valid conclusion" Responses.

Experiment	Figure: 1	2	3	4
Bara	3%	10%	33%	34%
Steedman1	11%	20%	21%	22%
Steedman2	5%	18%	19%	12%
Inder	17%	18%	25%	20%

Table 6.6: Performance of Subjects on Items Supporting Valid Conclusions.

feature of the problem or the way it is tackled. Different groups of subjects are clearly influenced by figure in different ways.

Whatever view is taken regarding the role of capacity limitations in the effect of the number of models, performance errors are so central to Johnson-Laird's theories that he is almost hiding behind the skirts of the execution errors, rather than facing the real task of explaining illogical behaviour in a satisfactory manner. Three model problems are said to be close to the limit of undergraduate subjects' abilities, either because they are approaching limit of "working memory", or because they require an amount of processing that makes random performance errors very likely (80% for Bara's subjects). In either case, subjects' performance can only be expected to degrade completely with any significant increase in problem complexity. If the mechanism for solving syllogistic problems is in any way relevant to everyday life, this makes it seem unlikely that anyone can ever manage to play an average board game, let alone understand simple hypothetical arguments or the publications of the Inland Revenue!

The problem arises from the fact that the model is built to explain a batch of data taken from a group of individuals. As a group, the subjects exhibit a spread of behaviour. Johnson-Laird, along with almost every other theorist in the field, tacitly assumes that the group is totally homogeneous and that each individual has exactly the same approach to the problem, and will therefore exhibit the same spread of answers. However, this is an act of faith, which Newell has counseled against:

> Much current work in syllogistic reasoning proceeds by positing a particular representation and / or method... These studies commit the *fixed method fallacy*, which attributes to all subjects (on all occasions) the same method without any attempt to ascertain either that all subjects indeed follow the specified method or that the consequences derived are invariant over different possible methods. They take as indicative of the model's success that it generates results in reasonable agreement with group data on percent correct.
>
> Many of these results, however, are not unique to the method but flow from disparate methods defined for different problem spaces.
>
> (Newell, 1981. P710)

Having made this basic assumption, and driven by a need to allow the potential for rationality, Johnson-Laird has designed a mechanism which is capable of giving the correct answer to each syllogism, and installed this in the head of each of his subjects. Incorrect answers are then seen as due to malfunctions of this machine – performance errors. To get the observed range of responses, the mechanism has to be made prone to stopping prematurely and spitting out incomplete results. These incomplete results do indeed correlate well with the observed spread of responses, although this can be seen to be a result of the fact that the notations that Johnson-Laird uses capture well the factors relevant to the task – the problem is being "carved along the joints". Moreover, the price for this correlation is exorbitant: Subjects are depicted as being unable to think about a fairly simple problem for as long as a minute without falling victim to a performance error.

6.6. Summary

Any theory of syllogistic reasoning based on mental models benefits from being able to offer a natural explanation for the observations surrounding illicit conversion. However, Johnson-Laird's particular brand of theorising entails four extensions to the obvious ideas of mental models – arbitrary numbers, uncertainty, negation and directionality – which conflict with the very features that distinguish mental models from propositional accounts. In addition, neither they nor the procedures for manipulating them are obviously applicable beyond syllogistic reasoning problems, although Johnson-Laird clearly believes they should be. Moreover, the manipulation procedures predict error patterns which have only variable empirical support, and sweep all the detail of subjects' performance, together with the connection of syllogistic reasoning with everyday mental activity, under the rug of "performance errors".

CHAPTER 7

A New Theory of Syllogistic Reasoning

7.1. Outline of the Proposed Theory

The previous chapter has suggested that there are a number of problems and uncertainties with Johnson-Laird's use of mental models to explain the way untrained subjects solve syllogistic reasoning problems. Of course, it is (almost) always possible to extend or modify a theory to take account of any particular (set of) criticisms or explain any new set of observations. For instance, Lee (1983a) captures how a workshop at the School of Epistemics in Edinburgh demonstrated conclusively that there is enormous scope for amending, extending or re-interpreting Johnson-Laird's proposed range of model manipulation procedures Unfortunately, so extending a theory in the absence of positive guidance from the original author – in the form of a clearly stated paradigm or cognitive architecture[1]

– quickly leads to a host of possible modifications, many of which are mutually inconsistent. Moreover, the process inevitably reaches that ill-defined point when the changes have to be regarded as constituting not an extension of the existing theory, but an entirely new one based on it. With Johnson-Laird's account of syllogistic reasoning, the contentious features are so central that this point is clearly not far distant. Therefore, rather than attempting to "clarify", "re-interpret" or "extend" it, an entirely new approach will be presented which more obviously meets the criteria for a good theory while remaining true to the spirit of his enterprise.

The new account shares with Johnson-Laird's the assignment of a central role to a kind of representation of situations – mental models – that is assumed to be important throughout a range of psychological phenomena. Moreover, it is able to account for syllogistic reasoning without needing to extend the power of these representations – i.e., it is able to do without conceptual mental models. As a result, it is able to avoid increasing the flexibility of the cognitive architecture that supports and manipulates the models, and thus in proportion weakening the theories that postulate them. It was inspired by the theory of Johnson-Laird and Steedman (1978), and attempts to preserve its impressive empirical accuracy and valuable insights. But although it bears an undeniable resemblance to its predecessor, the new proposal is based on a fundamentally different view of the role played in everyday life by the processes involved in reasoning experiments. While both theories assume that the mechanisms of everyday thinking play a central part in syllogism

[1] Rather that just one or two possible modifications, such as his "straw man" approach that was discussed at the end of Section 6.5.2.

solving, Johnson-Laird also believes that subjects' experimental performance gives an indication of the operation of that mechanism. The new theory, in contrast, stresses that although the underlying mechanism for information combination might be the same, its behaviour is obscured by the other aspects of the task, the difficulty of which is grossly underestimated or mischaracterised by Johnson-Laird.

The essence of everyday reasoning is that it does **not** involve dealing with "nonsense" premises – sentences that are not really "about" anything. Everyday thinking, "rational" behaviour and even survival crucially depend on the integration of new information with what is already known. Keeping it and manipulating it in isolation is a specific skill – it has to be taught as the foundation of formal logic, and acquiring it is difficult. Similarly, explicitly drawing a conclusion is a very uncommon thing to do. There is no doubt that the consequences of new knowledge are continually recognised and acted upon, but if the idea of mental models is correct, this can be done without making them explicit. Certainly they are very rarely enunciated: indeed, quite the opposite. If something is necessarily a conclusion of two explicitly known premises, it can be assumed to be generally recognised as such, and the Gricean objective of being informative argues against pointlessly stating the obvious.

The processes of everyday thinking are supported by a cognitive architecture that is genetically specified and has been exercised and refined upon vast quantities of common experience. As a result they can be expected to be stable and uniform throughout the population – at least to the extent that different subjects will be "similar". Johnson-Laird (in common with the majority of other theorists in the area) believes that syllogistic reasoning experiments reflect those processes, and thus that the results of groups of subjects will tend to reveal the shared parameters of the mechanism. This leads him (and them) to focus on the trends observable within groups of subjects, to postulate a mechanism which is basically geared towards producing sound reasoning, and to attempt to characterise reasoning errors in terms of its malfunctioning.

The recognition of syllogistic reasoning as an unnatural, formal, task leads to a considerable shift in emphasis. A subject's performance during the experiments reflects the response of an individual to an unfamiliar problem. Thus, there is no reason to expect that the strategies they employ should in principle be capable of "correct" performance, although the vexed question of the potential for sound reasoning cannot be ignored. Further, it seems likely that different strategies will be used by different subjects, or even the same subject at different stages of the exeriment – there is no reason to commit what Newell (1981) termed the *Fixed Method Fallacy*. As a result, the new account focuses closely on the responses of individual subjects, and suggests how they may be explained by the interaction of simple heuristics such as those that an untutored subject might be expected to apply.

Following Johnson-Laird's lead, and by way of conveying the relevant intuitions, the details of the theory will be introduced by way of the key "anecdotal observation" that inspired much of it. It occurred while trying to explain, in a social situation, what syllogistic reasoning was about. When a friend was given an example syllogism to do, she immediately asked for a pen and paper, and wrote down the premises. With them in front of her, she began to think about the problem,

using the pen to point to words or phrases within the two premises as she mouthed (presumably) the resulting sentences to herself. What was happening was obviously nothing like simply reading a potential conclusion off some internal model: it was clearly a difficult and time-consuming task. If, however, actually enunciating the conclusions of deductions is a very uncommon activity, as was suggested above, there is no reason to expect anyone to be particularly adept at doing it. This approach is also substantiated by the general reaction of people who are told of a conclusion that they have failed to find. They rarely have any difficulty in appreciating its validity: more often, they are **surprised**. This is not the reaction of someone who is overwhelmed by the logical complexity of the problem. It is the reaction of someone who had not thought of all the possible conclusions.

The original informal observation led to a further insight: the process of conclusion generation seemed to actively involve the lexical form of the premises rather than their logical properties. It seemed that potential conclusions were being generated by re-arranging the words of the original premises. Similar intuitions follow from informal observations of subjects attempting to solve syllogisms without having the premises written down. The process typically involves several heavily stressed and hesitant verbalisations of the premises, again suggesting that they are being fragmented and rearranged. Moreover, this should not be surprising. When presented with two sentences (premises) and told to offer another which is related to them, building the new from the same pieces as the old seems a very obvious way to proceed.

There is, of course, a trivially obvious way to get around the problem of conclusion generation. Since there are only eight types of possible conclusion, it is possible to simply try each in turn. However, this approach is unlikely to be adopted. It is quite out of line with the typically haphazard way in which, as Johnson-Laird (1983, P45) points out, most people approach reasoning problems. Further, realising the limited number of responses, and enumerating and learning them, would require thinking not about the particular problem presented, but about the family of problems and their solution. However, there is a strong sense in which this kind of meta-reasoning is "cheating" – subjects are told, and seem universally to accept, that they are to solve the problems presented to them as unrelated problems. This is also born out by the fact that none of Inder's experimental subjects asked how many problems there would be in the test, which would be very useful in trying to analyse the structure of the problem domain.

Finally, it may appear that this observation conflicts with the widespread intuition of those who had taken part in Inder's experiment. As reported in Section 6.2.2.10, subjects almost universally reported using (some variant or corruption of) Euler Circles to help them solve the syllogisms. However, this is not the case, and the way the two can be reconciled, and indeed seen as complementary, will be fully discussed in Section 7.4. For now it is sufficient to observe that while the surface forms of the premises are postulated to have an important effect on the generation of possible conclusions, this is far from the whole of the process of finding valid conclusions to syllogisms.

Inspired by these observations, the new theory recognises that the generation of conclusions is a difficult task, typically involving the surface properties of the premises. It agrees with Johnson-

Laird that syllogism solving involves the generation of a range of possible conclusions and then testing them to reject those that are inappropriate. Mental models are still central, but are now used to validate or test the conclusions once they are generated. When they are used in this way, there is no need for them to constrain or explain any features (other than the content!) of the conclusions that are offered. This means that there is no need for them to be directional or to have explicit negation, and the most objectional features of conceptual models can be eliminated.

This view of the process is very literally an instantiation of a *generate-and-test* process which Newell and Simon (1972) have argued is one of the most general ways of doing things. However, this very generality also strictly limits the ability of the approach to take advantage of the structure of any specific task, which leads Newell and Simon to reject it as too weak to account for skilled behaviour. However, this is not a problem in the context of syllogistic reasoning if it is recognised that for most people it is a strange and unfamiliar task, and the way they go about can hardly be regarded as skilled. Instead, the generality of the method makes it a very likely choice for a new, unfamiliar situation.

The essence of the theory can be presented as suggesting that syllogism solving is a four stage process:

(1) Subjects understand and internally represent the premises, by a process of building a mental model of the situation (or some sort of idealisation of an example situation) that the premises describe.[2] This utilises many of their ordinary linguistic abilities, although they will interpret the sentences somewhat abnormally because they know that they are in some sense "nonsense" – they are not "about" anything. In the absence of any context or semantic cuing from the content words, each kind of premise will be interpreted as indicating a single, "typical" situation. However, any models produced will be in direct correspondence with single (possible) situations, and not any kind of encapsulation of the truth conditions of the premise.

(2) Subjects attempt to generate a candidate conclusion from (something very close to) the surface form of the premises. To do this, they employ a selection of heuristics, such as re-sequencing the words as they appear in the premises or "splicing" some kind of internal representation of the surface (perceptual) properties of the premises. The objective of this stage is merely to produce something of the right form to be a conclusion, and while sequences of words might be rejected or morphologically modified so that only syntactically acceptable sentences produced, the logical properties of these sentence are **not** considered.

As suggested above, people unfamiliar with syllogistic reasoning will not generate candidate solutions methodically but will typically employ one or two strategies for producing a conclusion, possibly shaped by of their encounters with quantified sentences in everyday life. However, this is not to say that the search is **necessarily** haphazard – much of the skill of syllogism solving is an ability to quickly select relevant possible solutions, and its perfection

[2] Once again, there is an apparent discrepancy with subjects' intuitions about the use of Euler Circles. This time the conflict is more direct, and once again Section 7.4 discusses how it can be resolved by appealing to the properties of the restricted mental models being employed.

would be the rote recall of the appropriate conclusion.

(3) Once a candidate solution is produced, subjects will submit it to a sequence of tests. In particular, they will check whether it is consistent with (i.e. true in) the model created from the premises. If it is false, then clearly it is not a valid conclusion (since the model represents a counter-example) and it is rejected. However, if the existing model supports the conclusion, the subject may actively attempt to falsify it by considering other situations (mental models) consistent with the premises, generated using a set of model-modification heuristics. If they succeed in producing a model in which the tentative conclusion is false while the premises are both true, then once again it is rejected.

As with conclusion generation, conclusion testing (i.e. exhaustively considering the range of possible situations consistent with a pair of premises) is not a normal activity, so naive subject's procedures will be disorganised and inefficient. Once again, though, this need not be true, and the skilled syllogism solver will have developed reliable and efficient means of recognising how best to attempt to defeat any particular conclusion.

(4) If a subject cannot disprove a potential conclusion by any of the model manipulation techniques that seem suitable, s/he will accept it as valid. On the other hand, if it is rejected because a counter-model has been found, the conclusion-generation heuristics will be employed once more in an attempt to produce another conclusion, and the whole cycle repeated. If ever the conclusion generation heuristics cannot suggest a possible conclusion, the subject will respond that no valid conclusion can be drawn (at least, by him/her).

This presentation of the new theory is like Johnson-Laird's account in that it tends to suggest a single generate-and-test cycle, using only a single (logical) kind of test. However, this is an over-simplification. A satisfactory conclusion must satisfy constraints of many different natures – some of these are linguistic (it must be a sentence), others relate to its logical properties (it must be true in all possible models) while still others are pragmatic features of the experiment (it must be in one of the syllogistic moods and it must link the end terms). It must pass at least three different types of test – syntactic, pragmatic and semantic. Johnson-Laird circumvents this problem by ensuring that the only things that can be generated – read from the model – are necessarily syntactically and pragmatically acceptable, and thus he has only one stage of (semantic) testing (and therefore only one theoretically significant locus for variability of behaviour). However, within the new theory each of these types of testing is recognised as an explicit stage, requiring skills that are not common in everyday life and allowing the opportunity for error.

This suggests that the new theory should be thought of as a generate-and-sequence-of-tests loop, where if any of the tests in the loop is failed, the cycle is repeated. However, this same flow of control can be conceptualised a different way, as a nested set of generate-and-test loops. What appears at any particular level to be a simple generator might be just that, a procedure which constructs the sequence of candidate conclusions that are observed. Alternatively, it might itself be analysed into a complete generate-and-test system, with an over-general candidate producer and a suitable (set of) filters. Thus many details of the subjects mental processing are impossible to determine from their responses alone. For instance, response distributions cannot determine

whether subjects simply only generate (consider) sentences in the syllogistic forms, or whether they generate sentences of any kind and then rejected those that were not of the appropriate form. Of course, the arguments against Anderson (1978) advanced by Pylyshyn (1979b) and cited in Section 0.4 suggest that tackling this indeterminacy involves taking account of more than just the generated responses. In particular, response times seem a very obvious feature to consider. However, the situation is not straightforward, as will be suggested when the matter is raised again in 7.4.

The essential feature of the new account is that an individual's unsound reasoning is not ascribed to any capacity limitations or slips (execution errors), but to weaknesses in the sets of heuristics that are brought to bear on the tasks of generating and falsifying conclusions. In particular, the theory proposes that syllogism solving behaviour arises from the interaction of a few simple rules which can themselves be justified or are at least plausible on other grounds, This has a clear similarity with the basic activity of linguistics, where building a grammar is very much a matter of finding sets of simple elements (transformations and filters, or categories for words) which combine to produce the observed range of language behaviour. However, very little has yet been said about the kinds of heuristics that subjects are postulated to employ, a matter which is clearly central to the account and thus obviously in need of attention.

Because of the number of different syllogisms, the complexity of the patterns in a subject's responses and the subtly of the interactions between different solution strategies, it is almost impossible to tell from the rules alone whether any particular response can be explained. For this reason, a computer program has been used to "animate" sets of solution strategies, and thus clarify the extent of the fit with the actual observations. The next section describes the program that has been used for this, while Section 7.3 goes on to present a number models that have been built to capture the behaviour of specific subjects. In doing this, each section will offer examples of the kinds of strategies that subjects are hypothesized use for generating and defeating possible conclusions, while section 7.4, which contains a more detailed discussion of the overall theory, offers some generalisations about them.

7.2. Using a Computer Program to Support the Theory

At the end of the previous section it was suggested that the motivation for the use of the computer program was to make clear the consequences of proposing any particular combination of heuristics, in terms of responses the subject would be predicted to have made. However, there is another aspect of the use of computer models in cognitive theorising.

In the introduction to this thesis it was suggested that, given our current tools for describing processes, Cognitive Science should be seeking to put forward an account of mental activity in terms of computational information processing. This raises the issue of the extent to which the methods used by a computer program which produces the same behaviour as the subject should be interpreted as relevant to those employed by that subject. In Pylyshyn's terminology, it is important to consider the extent to which the program can be seen as more than just weakly equivalent to the subject. Unfortunately, computers and brains are utterly dissimilar computing engines, and there is no reason to imagine that the functional architecture supported by modern programming languages bears any resemblance whatsoever to the cognitive architecture that

underlies mental phenomena. This means that the interpretation of the machinations of a computer program as relevant to human mental processes is fraught with potential problems. As a result, before the details of the simulation program are described (in Section 7.2.2), the next section will offer a brief discussion of how it is meant to be interpreted.

7.2.1. Programs and Psychological Models

Whatever theory of mind one adopts, it is possible to attempt to produce computer programs that exhibit *weak equivalence* to a piece of mental activity, in the sense of offering the same responses to the same stimuli. However, the view of Cognitive Science outlined in Chapter 1 assumes that mental processes arise from computations carried out within the brain. This opens up the possibility that a computer program can achieve *strong* equivalence, where it could be seen carrying out the very same computation that occurred within the brain.

Because every computation, including those within the brain, takes place within a computational architecture, part of the commitment in proposing a strongly equivalent computational model is to a particular architecture. Given what is known about the brain, the cognitive architecture is not that of any existing electronic computer. This means that

> Modelling cognitive processes must proceed in two phases. The first is to emulate the functional architecture of the mind, and the second is to *execute* the hypothetical cognitive algorithm on it (not to *simulate* the behaviour of the cognitive process, for the algorithm represents a literal proposal for the structure of this process.
>
> (Pylyshyn, 1980. P120)

It must be recognised that any computational model that seeks strong equivalence with a psychological process must identify and emulate an assumed cognitive architecture. This means that its psychological significance can only be understood if the proposed architecture can be identified, and to the extent it is possible to discern which features of the simulation are relevant to the simulated.

In many cases the presentation of the model will clarify what the theorist takes to be the distinction between the parts of the model that emulate the cognitive architecture and the parts that execute the proposed psychological algorithm. Typically, the latter will be exhibited while the former will only be characterised. For instance, Marcus (1977) illustrates how some phenomena observed in human linguistic behaviour follow naturally from the operation of a specific (complex) mechanism or cognitive architecture. The rule packets themselves are listed in full, while the stacks and windows of the parsing machine are given only a functional characterisation. Similarly Brown and Burton (1978) and O'Shea and Young (1977) present subtraction skills in terms of (specific, exhibited) rules suited to a cognitive architecture based on a production system (see Newell and Simon, 1972), which Young (1973) also applies to children's seriation behaviour.

However, this distinction is not universal. Luger, Bower and Wishart (1983) still appear to attach great importance to a program, which they describe in considerable detail, which is intended to relate to some features of infant gaze behaviour when tracking objects. However, closer analysis (Inder, 1983) suggests that this program cannot be interpreted in terms of the execution of a plausible psychologically instantiated algorithm on an emulated architecture, and concludes that its worth is severely limited by this.

It may seem possible to try and steer a middle road, proposing mechanisms that are "fairly equivalent" – that is, more than weakly, but less than strongly. Proposing such a situation might result from an attempt to capture more than just the superficial features of the mental behaviour while admitting inability to specify the mechanisms precisely. For instance, a theorist who firmly believed the brain to be highly parallel might nonetheless offer an account couched in terms of sequential processes. In fact, once it is recognised that human limitations may thwart laudable ambitions, Pylyshyn's sharp distinction remains, though appearing more applicable to the intention of the theory than its achievement. Theories must strive for strong equivalence, though seldom actually expecting to achieve it, whether for reasons of empirical inaccuracy or (as in the parallelism case) of deficiency of conceptual tools.

Limitations in the scope of a theory are another reason for discussing a model (and thus advancing a theory) despite believing that it does not accurately reflect any mental process. This applies to the details of the current theory (and accompanying model) of syllogistic reasoning, where the behaviour of any subject can be captured in a number of ways (i.e. is underdetermined by the data available). In this case there are a number of other factors, over and above distribution of responses, that are relevant to determining the choice of the actual strategies that any subject is postulated to have employed. Foremost amongst these is a domain-independent theory framework in which to set the account. In the case of the proposed theory, this is a commitment to the use of a mental model as the focal information structure, which serves to impose upon the account a number of limitations on what can be represented, but places no constraints on the strategies that may be deployed. In particular, it says nothing about how they develop during the experiment, and thus the program simply models them as appearing fully formed in a convenient form. Nevertheless, it remains potentially a literal account of the computations underlying a particular individual's syllogistic reasoning performance, at least at some level of detail. It is offered in the explicit knowledge that it is probably wrong, although in this it is (or at least should be) no worse than any other scientific theory (See Chomsky, 1980: Chapter 1).

7.2.2. The Syllogism Solving Program

The program written to illustrate and support the current theory attempts to generate valid conclusions to syllogistic problems. In doing so, its purpose is to make explicit the conclusions that would be expected on the assumption that the problems were being tackled in a specified manner. To do this, the program allows the skills or procedures being evaluated to be encoded in the form of a set of conclusion generation and falsification strategies. It then allows the user to type in a pair of syllogistic premises and produces the conclusion that would result from the application of those strategies within the overall approach outlined by the theory presented above. Alternatively, the program is able to process batches of syllogisms and produce a table of the responses that would be given by those strategies. When working in this mode, the program actually accepts syllogism specifications in the same format as that used by the system used to present them for the experiment. This allows it to work directly from the files specifying the problems actually used to test the subject. The user is also able to instruct the program to modify the group of strategies it uses as it progresses through the batch, which allows it to reflect the effect

of the subject learning during the experiment.

The program was written in Prolog,[3] which was chosen because it is easily able to express pattern-matching operations, and the basic operation most central to the theory of syllogistic reasoning is the pattern based selection (and subsequent application) of solution strategies. It does have drawbacks, particularly related to the clear expression of sequences of activities, since the relation between the "flow of control" and the structure of the code is not intuitive. Poplog[4] would have been better suited in this respect, since it would have allowed the explicit expression of the sequential parts of the control flow in a procedural language. However, the Poplog system available when the program was being developed was unacceptably slow, particularly because, since the program sought to clearly express specific strategies, it could not be coded for efficiency.

The program consists of about 1000 lines of sparsely commented Prolog code, with a further 200 lines or so potentially specific to capturing each individual subject. It takes between 6 and 10 seconds of CPU time on a VAX 11/750 to process each syllogism. This is between 10 seconds and a minute of real time per syllogism, or between fifteen minutes and an hour and a half to produce a conclusion for an entire batch of sixty-four. From the viewpoint of the theory of syllogism solving being presented, it can be broken down into four main sub-sections:

(1) **Model Construction and Manipulation:** The theory assumes that mental processing, and syllogism solving in particular, makes use of the construction, manipulation and examination of mental models. The simulation of this process involves data structures that allow entities and their properties to be represented. In addition there are procedures for creating entities, adding properties to them, checking what properties they have and selecting those with particular properties. Together these data structures and the procedures for accessing them simulate the postulated manipulation of a mental model by the cognitive architecture.

(2) **Processing Language:** The theory assumes that in solving syllogistic problems, subjects make use of their normal procedures for handling language, and that these involve the use of mental models. The program has procedures for accurately checking the veracity of sentences (in syllogistic form) against the state of the model and others for modifying the model as necessary in order to make such sentences hold. In their operation, these procedures call upon those others, just mentioned, that support model construction and manipulation.

(3) **Syllogistic Reasoning:** The program must obviously contain procedures for actually solving syllogisms – that is, for co-ordinating the linguistic processing of the premises and the selection and evaluation of potential conclusions. The proposed theory suggests that the way subjects perform when solving meaningless categorial syllogisms depends on the details of the ways in which they generate and test potential conclusions. Each of these activities is assumed to be an explicit skill, which the program allows to be captured as collections of strategies (which take the form of condition-response pairs) that are specific to each subject.

---

[3] The language is fully described in (Clocksin and Mellish, 1981).

[4] A combined programming system allowing Prolog to be combined with Lisp and Pop: See Sloman and Hardy (1983).

Thus the program allows one to specify the way a subject tackles syllogistic reasoning in terms of collections of strategies for conclusion generation and conclusion falsification. These are then applied within the framework of an overall strategy which is constant for all subjects.

(4) **Control and I/O:** Finally, the program contains procedures for such tasks as reading from files or terminals and printing or tabulating responses and other diagnostic information. There are also procedures for controlling the program through ordered batches of syllogisms. Moreover, since the theory assumes that the skills involved in syllogistic reasoning are unfamiliar, the program also allows one to specify not only the sets of strategies to be used, but also the way they are to change during the solution of a batch of items. Thus it has procedures which can modify the sets of strategies that the syllogistic reasoning employs in pre-determined ways, although no significance should be attached to the manner in which this is accomplished.

All but the last of these, which embodies only the requirements of making an operational program, will now be presented in greater depth. Then an example of the way the program simulates the solution of a particular syllogism will be discussed. Finally, some general comments on the limitations of the program will be offered.

7.2.2.1. Model Construction and Manipulation

At the core of the program is the production of behaviour reflecting the construction and manipulation of models of situations. These models take the form of information structures which specify the features of the things being modelled. Specifically, each entity in a model is described by a Prolog ground clause which expresses a number of attribute/value pairings for the entity. For instance, the clause

entity(entity_34, props([prop(colour, green)])).

could be used to indicate that green was the *value* of the *attribute* colour of the *entity* "entity_34". Similarly, of course,

entity(entity_34, props([prop(colour, green), prop(shape, sphere)])).

would suggest that it was a also sphere. It should be stressed that, since the properties are gathered together in a list, there is an ordering between their representation. However, this ordering does not signify anything about the properties themselves, which are assumed to be independent (or at least unordered). In fact, the model accessing procedures pay no attention to the ordering of the properties within the list, and thus it exerts no influence on the course of reasoning.

It is important to realise that although this example suggests that the program represents attribute values in familiar ways – using words like "green" or "sphere" – this is not intended to reflect any feature of the models constructed by the cognitive system. Indeed, it seems far more likely that colours will be encoded as some continuously variable indication of the relative stimulation of retinal cell types, or that shapes may be captured in terms of collections of oriented surfaces. There is no reason to suggest that the content of a mental model must necessarily be captured or structured in terms of lexical items from any human language. This is, of course, the

characteristic feature of the last of the three possible interpretations of Johnson-Laird's models that were outlined in Section 6.5.1, which was also suggested to be the most true to his aims.

In accordance with this view, the program provides mechanisms whereby the modifications to the model that are triggered by the use of a word do not involve the word itself in any form. In particular, the program implements a mapping inspired by the suggestion that the meanings of words are closely associated with prototypes. It allows common nouns (the only open word class it handles) to be defined using object templates. Each noun can be associated with a collection of attribute/value pairs which characterise members of that class.

This is illustrated by the following model:

```
entity(1, props([
        prop(wearing, silly_hat),
        prop(holding, palette)
        lprop(is_a, [syllogism_entity])])).
entity(2, props([
        prop(wearing, silly_hat),
        lprop(is_a, [syllogism_entity])])).
```

It depicts two syllogism entities, both of whom are "wearing a silly hat" and one of which is also "carrying a palette".[5] This model is somewhat similar to what might have been built by the program in response to the premise "some of the beekeepers are artists", assuming the words involved were associated with the following prototypes:

```
prototype(beekeeper, props([prop(wearing, silly_hat)]))
prototype(artist, props([prop(holding, palette)]))
```

The properties used to characterise these prototypes were chosen to emphasise that they are not to be thought of as reflecting psychologically plausible dimensions along which such entities might actually be represented. Moreover, the fact that these definitions specify only a single property as associated with the word is misleading. Within a system capable of capturing meanings to any significant extent, the prototypes associated with words would specify (ranges of) values for a number of factors, none of which is independently significant. Certainly no single attribute value can be expected to capture any kind of definition of the word, and as a result there will typically be no single component of the representation associated with the property that can be ordered or negated.

To support this style of representing word meanings, the program provides a primitive model-access procedure which is able to determine, on the basis of these templates, whether a word is applicable to any particular entity in the model. There is also another, closely related procedure which attempts to amend its representation to make it fit. In doing so it will necessarily encapsulate within the representation of the entity certain aspects of the words semantics – namely its characterisation along a set of dimensions which are independent of any lexical item. As a

[5] Notice that one of the properties of the entity – the fact that it "is_a" syllogism entity – is represented differently. This is the way the program allows multi-valued properties to be represented. However, it does not reveal a commitment to their relevance to psychology and such properties are possibly best thought of as shorthand for complex and non-interfering patterns of other properties. Since the multiple values for such attributes are represented on a list, they are, in fact, always ordered. However, as with the normal properties themselves, this ordering is not allowed to influence the results of the program in any way.

result, the data structures built by the program, together with the procedures that manipulate them, exhibit the properties of a mental model.

These model-access procedures within the program work on the basis that a noun can only be applied to a modelled entity if for every attribute associated with its definition, the entity has a value which matches the value in the template. A more sophisticated system could allow partial mis-matches on some attributes, or even distinguish between essential and usual features, possibly on the basis of context. Doing so would allow it to offer a measure of the degree to which the term applied to the object. Thus it could produce a quantitative "uncertain" response even though all the information structures involved are taken to be definite. However, the program only implements the simplistic all-or-nothing approach, which is adequate for categorial syllogisms.

The decision that a template can only be deemed to fit an object if it matches in every respect heightens the significance of any change to the model. Under the kind of partial matching system just mentioned, changing only a proportion of the attributes of an object that are specified in a template would degrade, but not necessarily destroy, the fit between them – the object would be represented as a less-typical exemplar of the class. Such continual refinement of a representation would be the normal way of using a non-lexicalised model for interpreting a description in a discourse. However the program's simplistic requirement for a complete and exact match means that if two templates specify different values for any attribute, they will match mutually exclusive sets of entities – i.e. they correspond to contradictory words.

In a situation where making an entity fit a template requires changing a value, doing so would have the effect of imposing the new classification in place of the previous one. The entity cannot be represented as being in both categories, and it is necessary to choose whether to remove it from its former category (by changing the value) or to abandon the attempt to make it fit the new template. On grounds of both computational efficiency and intuition, the procedure for making a model fit a template consistently follows the latter policy. It is sensitive to the distinction between supplying and modifying a value for an attribute, and while it will extend the representation of an entity by adding values for a previously unspecified attribute, it will not change any value that is already specified. If it is called upon to do so, it will give up and indicate that it cannot.

These mechanisms for handling word meanings have served to illustrate the kind of way in which the semantics of the lexical material can influence the behaviour of a model-based computer program, even thought the words are not present in the model. However, they have played no part in reproducing the behaviour of particular subjects reported in section 7.3. No definitions were provided for any of the words involved in the syllogism modelling, and in such circumstances, the program defaults to simply adding undefined lexical items to a list associated with the each relevant entity. For instance, if the second entity in the example above were to be ascribed some property for which the program had no template, such as "stamp collector", it would be modelled thus:

```
entity(2, props([
        prop(wearing, silly_hat),
        lprop(is_a, [syllogism_entity, stamp_collector])])).
```

This corresponds to simply noting a connection between the entity and an uninterpreted word. Such a representation cannot support any kind of implicit recognition of class inclusions or

contradictions which would be possible if words had full definitions. However, it is adequate for modelling the performance of subjects during the experiment, provided one assumes that the lexicalisation of the experimental items did not influence them.

Finally, and still in line with the preferred interpretation of Johnson-Laird's models, the program does not have any facilities for indicating or restricting the context in which the value is applicable to the attribute of the entity. That is, there is no mechanism for subdividing a model or simultaneously entertaining more than one. Within a model, neither attributes nor values can be ordered, and there are no mechanisms for giving partial specifications of an attribute value. In particular, it is not possible to exclude a particular value for an attribute – to indicate that it does **not** have some specific value.

7.2.2.2. Processing Language

As well as being able to determine the applicability of any word to an entity, subjects are assumed to be able to determine (i.e. the computer program provides routines for determining) whether sentences in each of the four premise forms are true in the current model. Within the program, determining the applicability of an assertion of the form "Q X are Y" is a three-stage process. Initially the set of all entities in the model which can be described by the noun X (i.e. those which satisfy the template associated with X) is formed. Then the corresponding set associated with the noun Y is formed. Finally, the veracity of the assertion is evaluated using a set operation appropriate to Q, the quantifier of the sentence, and the presence or otherwise of a negation (i.e. the word "not").

The fact that the program employs this particular algorithm is not intended to be psychologically significant. The claim is merely that, as part of their ability to speak English, people have some method by which they are able to correctly evaluate whether a sentence in syllogistic form is true in a particular mental model. Moreover, since this is a facet of knowing the language and the meaning of the quantifiers involved, every subject is assumed to possess established (i.e. stable) methods for performing this evaluation accurately and reliably. As a result, all subjects have been modelled as using a single unchanging procedure at all times.

In addition to allowing the evaluation of a sentence in syllogistic form against a model, the program also provides the means for extending or modifying models to **make** such sentences hold. The two tasks have much in common and are tackled in essentially the same way (in fact, the program ultimately employs a single parameterised procedure for both tasks), gathering the relevant sets and evaluating the relation between them. If the relevant set relationship does not already hold in the model, the procedure for making premises true will attempt to use the model modification procedures mentioned above to add properties to entities in order to make it hold. However, where this would involve dealing with an empty set, it first creates new entities to which the relevant properties can be added. As a result, the program incorporates the Gricean assumption that the problem only mentions properties that are exemplified.

Whenever the program has to introduce new entities to the model, it has to decide how many it should create. This decision is arbitrary, in the sense that there are no a priori factors to constrain it (other than the need to use more than one in order to facilitate subsequently

distinguishing between "some" and "all" of the group). There are arguments for choosing to create small numbers of individuals based on the effort of manipulating them. These are obviously relevant to the computations involved in simulating the process, and arguably also have intuitive psychological plausibility, since people would be unlikely to try to manipulate more entities than seem to be required. All the subjects models reported in section 7.3 make use of a group size of 2, and informal exploration suggests that this parameter does not have a great effect on the responses produced.

As with the verification procedures, there is no claim made about the precise way in which subjects may carry out this process. Moreover it was argued in Chapter 3 that in the normal use of language this task would be heavily influenced by the subjects understanding of situation under discussion. When there is no sensible situation, as with syllogistic reasoning with "familiar nonsense" premises, subjects will be without an important factor guiding their normal linguistic processing. As a result, it quite is possible that the way the task is tackled may vary both between subjects or over time in the same subject. However, constructing a model in line with a sentence, like determining its veracity, is clearly closely tied to knowledge of language. As a result, although subjects may possibly vary in the precise models they create, they can nonetheless all be expected to always produce a model in which all (i.e. both) the premises hold. Despite the fact that the details of the model construction process constitute a potential degree of freedom of the model, it has not yet been found necessary to exploit it. All the subject models presented in the next section were produced using a single constant approach to premise interpretation.

When there is no current model – i.e. for the first of a pair of premises – using a constant procedure to capture the meaning of a sentence always results in the construction of the model for each premise type. By default, particular premises are assumed to imply set overlap – i.e. **I** and **O** premises both result in models in which both are true – while **E** premises are taken to imply the existence of disjoint (but non-empty) sets. Universal (**A**) premises are taken to imply a subset relationship, and thus result in a model which will not support the conversion of the premise. Using Johnson-Laird's notation, the models produced are:

A(x, y)		I(x, y)		E(x,y)	
	x = y		x		x
	x = y		and		x
	y		x = y		y
	y		O(x,y)		y
			y		y

The second premise of each pair is interpreted in the context of the model built on the basis of the first, and the model eventually produced arises from the interaction of the construction procedures with the structure of the existing model.

Note that there is no sense in which this process attempts to capture the truth conditions of the premises, and the resulting model will simply represent a situation that the subject thinks of as typically described by the sentence. Thus, while a subject may well realise (and verify) that "some a are not b" is true when no a are b, the model they build may well have some that are.

7.2.2.3. Syllogistic Reasoning

The overall strategy for the solution of a syllogistic problem is directed by a routine which is part of the subject-independent kernel of the system. It is responsible for controlling the invocation of the conclusion generation and falsification procedures.[6] The former are responsible for generating candidate conclusions from the surface form of (in the program, the lists of words in) the premises. The latter are responsible for attempting to defeat a conclusion by modifying the current mental model to create a situation in which it is false but the premises are true. It is the task of the top-level routine to coordinate these two activities, and ensure that any conclusion that the candidate generation procedures offer that cannot be defeated by the falsification procedures is taken to be an answer to the problem. Should a generated conclusion be falsified, the generation procedures will be given a chance to generate another, until they can no longer do so, at which point the problem will be assumed to have "no valid conclusion". Since the generating procedures actually used in modelling subjects are written to be methodical and so always give up gracefully, this approach is adequate. Were they not, or if the program were to be extended attempt to handle the temporal aspects of the process and model subjects working under a time limit (which it currently cannot) then there would obviously have to be some additional mechanism to cause the conclusion generation process to be abandoned. In this case, too, the item would be assumed to have no conclusion.

Unfortunately, the key feature of theoretical relevance in this top-level procedure is potentially misleading! It was suggested in Section 7.1 that the a subject selects conclusions on the basis of three distince types of criteria – syntactic, pragmatic and semantic.[7] Nevertheless, the program is structured as a single generate-and-(semantic)-test cycle, which might be taken to suggest that this is the only important type of testing. If fact, this stress is merely a consequence of the fact that the program was written at an early stage in the development of the theory. At that time, logical testing was recognised because of its role in Johnson-Laird's work, but no subject data was available to point out the importance of the others. The other tests are still applied, but this happens within the conclusion generation procedures – i.e. the generator in the overall generate-and-test cycle is actually itself a complete generate-and-test system. However, it was also pointed out in Section 7.1 that the position of any test within the hierarchy does not affect the range of responses that will finally be produced. Since the program makes no attempt to capture response times, this is of no significance.

Since it is simply a generate-and-test cycle, this level is uninteresting, but it is possible to delve more deeply into the details of the generation and falsification procedures. These both form part of the subject-specific portions of the code, and the operation of each will be described in turn.

---

[6] The Prolog code of this routine is shown in Fig. 7.1, where "nvc" is used to indicate the suggestion that there is no valid conclusion.

[7] I.e. are they sentences, syllogism forms and true statements.

```
solveit(Premises, Conclusion) :-
        generate_conclusion(Premises, Conclusion),
        not falsify_conclusion(Premises, Conclusion).
solveit(Premises, nvc).
```

The top-level control routine.

```
suggest_conclusion([ Premise_1, [all, Bs, are, Cs]], Conclusion) :-
        replace(Bs, Cs, Premise_1, Conclusion).

suggest_conclusion([ Premise_1, [no, Middle_term, are, Predicate]], Conclusion) :-
        replace(Middle_term, [not, Predicate], Premise_1, Conclusion).
```

Two conclusion generation strategies.

```
modify_strategy(Props,[no, A, are, B]) :-
        true([some, A, are, B],make).
```

A conclusion falsification strategy: true(Sentence, make) is the procedure to modify the model to make Sentence true.

Fig. 7.1: Examples of Prolog code from the simulation program
See text (Section 7.2.1.3) for an explanation of its operation.

7.2.2.3.1. Syllogistic Reasoning: Conclusion Generation

According to the proposed theory, the generation of possible conclusions for a particular premise pair is a specific skill. In the program this is manifested by modelling each subject as possessing an evolving collection of strategies for accomplishing the task. Moreover, both introspection and informal observation suggests that this process is based on surface features of the premises. In the program, this is reflected in the fact that the conclusion generation strategies pattern match against (and subsequently generate) lists of the words involved in the premises. It should be noted that while this is plausible, it is a falsifiable claim of the theory.

As suggested in the previous section, the overall conclusion-generation procedure is itself a generate-and-test system which invokes and monitors the collection of conclusion-suggesting procedures the subject is currently using. As such it has two main functions. The first, which has not actually been modified for any subject, is to methodically and exhaustively generate the different re-arrangements of the premises and the terms within them. Each permutation is then presented to the procedures that embody the conclusion-generating strategies that the specific subject employs. The second is to carry out the pragmatic and syntactic checks on each possible conclusion offered by the lower-level strategies. The syntactic checks – i.e. for sentencehood – are assumed to be part of the subject's knowledge of English, and have thus never been altered for (or in) any subject model. In contrast, the pragmatic checks are simply a feature of the experiment, and some subjects (despite the instructions) either never fully understand what is required, or only

come to do so as the experiment progresses. As a result, the pragmatic tests may well be different for each subject, or even change over time within one subject model (e.g. Subject B in section 7.3.2).

If a possible conclusion passes the tests that the subject is (currently) using, it is made available to the top level loop for the consideration of its semantic properties (i.e. truth in the model). If it is subsequently rejected as being falsifiable, or if it simply fails the pragmatic or syntactic tests, then, as in the top-level control loop, the generation procedure is given a chance to generate another. First any remaining conclusion-suggesting strategies are first re-invoked with the same arrangement of premises that gave rise to the unsatisfactory suggestion, after which all the strategies are tried with any as yet untried permutations of the premises. Thus this routine and the top-level controller together ensure that the program will only give up on a particular re-arrangement of the premises when every strategy has been considered for application, and will only declare that there is no conclusion when all possible arrangements have been so considered.

The lowest level conclusion generating processes in the program that claims anything other than the weakest equivalence to the subject being modelled are the actual solution generation strategies. These are the procedures which actually specify the possible conclusions that are worthy of considering on the basis of (some re-arrangement of) the premises. Such conclusion-suggesting routines are entirely syntactic mappings between the arrangements of words in the premises and an arrangement of words for a possible conclusion. They take no account of the semantic and pragmatic constraints on an acceptable conclusion. Within the simulation they take the form of very simple Prolog procedures (see Fig. 7.1), although of course this is of no psychological consequence. What is relevant is the kind of mapping that they specify, which is exemplified by the following strategy:

> IF the second premise has the form "all Bs are Cs"
> THEN a possible conclusion can be found by
> replacing any B in the first premise by C.

This is arguably about as obvious a strategy as one can get and, moreover, actually appears to be very common. The entire set of those actually used to capture the performance of the subjects presented in section 7.3 are outlined at the end of that section.

Note that even though the applicability condition of this example rule is very selective, the presence of the permutation process described above ensures that it will be matched against both orderings of the premises, and every permutation of terms within them. As a result, this one strategy will suggest conclusions for any syllogism involving an **A** premise, regardless of which premise it appeared in and whether the middle term was its subject or predicate. The same range of conclusions could, of course, be produced by applying this and three additional (and complementary) strategies only to the premises as presented. Doing this would also allow considerably greater flexibility, since the conclusions generated for each distinct type of **A** premise could be specified individually.

Unfortunately, there is no accepted framework within which to characterise the premises or the patterns within them or to delimit the possible range of manipulations of them when generating a conclusion. Thus the range of procedures used to model specific subjects is constrained only by

pre-theoretical notions of simplicity. In the particular case of the example rule, these criteria suggested that one strategy in combination with a universally applicable permutation process was more parsimonious than four independent strategies. Moreover, postulating processes that operate only with the premises arranged as presented seemed at odds with the informally observed behaviour of tracing round the premises in different orders (described in Section 7.1), which is highly suggestive of some kind of (apparently haphazard) permutation process. These issues are raised again in more detail in Section 7.4.

7.2.2.3.2. Syllogistic Reasoning: Falsifying Conclusions

The other procedure invoked by the top-level loop is responsible for attempting to modify the state of the mental model in order to falsify the conclusion currently under consideration. As with the generation of conclusions, the proposed theory suggests that this is also a task-specific skill, which is again reflected in the program by modelling the subject as possessing a number of strategies tackling the task. As a result, the highest-level falsification procedure resembles its counterpart in the generation process, in that it is responsible for cycling through the subject's battery of model-modification strategies and checking whether any of them actually creats a counter example.

Falsifying conclusions, like generating them, is a specific skill which is acquired by subjects as they become familiar with the syllogistic reasoning task. It too is captured as an evolving collection of strategies, each specifying an action appropriate to falsifying certain types of conclusions. Within the program, the formulation of any particular falsification strategy is couched in terms of both the proposed conclusion and the premises that gave rise to it, although no strategy actually used in modelling subjects is at all sensitive to the original premises.

Unfortunately, as with the generation of conclusions, there are no a priori constraints on what can constitute an acceptable falsification strategy. The form used for those employed in modelling subjects is illustrated by the strategy shown in Figure 7.1, which can be glossed as:

IF the conclusion to be defeated has
the form "No As are Bs"
THEN modify the model to make the sentence
"Some As are Bs" true.

In fact, all the model-modification strategies used to capture the behaviour of subjects have been formulated in terms of attempting to defeat possible conclusions by trying to make some other syllogistic-form sentence true. Thus all the subjects whose performance is discussed in section 7.3 have been modelled as having or acquiring some of the following strategies:

To defeat a conclusion of the form **A(x, y)**, make **O(x, y)** true.
To defeat a conclusion of the form **I(x, y)**, make **E(x, y)** true.
To defeat a conclusion of the form **E(x, y)**, make **I(x, y)** true.
To defeat a conclusion of the form **O(x, y)**, make **A(x, y)** true.

Subjects falsification strategies were couched in this way because it seemed to provide an adequate falsification mechanism without requiring any additional representational power and the minimum of task-specific procedures. It also makes the intention and basis of the strategy immediately clear, and is intuitively plausible since people seem well able to recognise such

contradictory sentences. Finally, by making the falsification process use the same representations and procedures as the initial interpretation of the premises, it imposes the maximum constraints on the overall system. However, since subjects are seldom very effective at falsifying conclusions, it is important to be clear that a failure to use a strategy like this does not by any means imply that the person does not know that the sentences were contradictory. What is missing is the realisation of the relevance of knowing how to generate a contradictory sentence to the task of defeating a conclusion. Moreover, it is perfectly possible that a subject will regularly use a strategy of testing a conclusion by trying to make a contradictory sentence true, and yet still not recognise the relevant generalisation. The knowledge to suggest doing so may well not be accessible to any kind of conscious processes.

Once the model has been manipulated by such a strategy, it is necessary to check that the possible solution has been invalidated. Since these strategies need (and typically) take no notice of the premises, this involves checking both that the conclusion has been falsified and that the premises either still hold or can be made true again. This decision – whether to allow further altering the modified model to make the premises true again – constitutes another degree of freedom of the model. The difference between these alternatives is subtle, and most subjects are modelled as adhering to the former.

7.2.2.4. An Example of Solving a Syllogism

To illustrate the way the procedures described above operate, this section will present a detailed description of the steps the program goes through when solving a particular syllogism. The problem on which the discussion will focus is an example of the **2AE** syllogism:

all beekeepers are artists
no chemists are beekeepers.

The subject model which will be assumed for this discussion will have only one strategy each for generating and falsifying conclusions. These are the strategies used to illustrate the discussion in Section 7.2.2.3 – namely the conclusion generation strategy associated with Lee's A-Effect, and the strategy for attempting to falsify **E(x, y)** by making **I(x, y)** true. These strategies will be applied within the overall framework of premise interpretation and validation procedures and control algorithms used as the basis for all the subject models to be presented in the Section 7.3.

The first stage in finding a conclusion to a syllogism problem obviously involves reading the premises and interpreting them, which under the current theory involves the construction of an appropriate model. When the program is interactively working with single problems, these are typed by the user as normal English sentences. As soon as each sentence is read, it is used to enhance the model that has been constructed so far. It turns out that for the first premise there happens to be nothing to enhance, since the model is cleared out at the start of each problem, but this makes no difference to the steps involved in interpreting the premise. When the program is working through an entire batch of problems, it accepts them encoded in the compact form used by the experiment presentation and analysis programs. This notation is then internally expanded into the corresponding sentences, which are then interpreted in order, just as if they had been typed by the user. In any case, once the second premise has been interpreted, the premise interpretation

process is repeated until no changes are required to make the premises true. This ensures that the model that has been created captures a situation in which both premises hold.[8]

The first step in solving the problem being described, therefore, is to construct a model for the first premise, "all beekeepers are artists". As described above, this starts by attempting to find the set of all beekeepers in the model. However, since the first premise is interpreted with respect to an empty model, there are none to find. This causes the program to extent the model by creating two new entities and making them instances of beekeepers – a strategy which reflects the Gricean principle (discussed in Section 4.2.2 and towards the end of Section 6.1) that types of entities would not be mentioned if they were not represented. If the program were interrupted at this stage, it would describe the model thus:

Entity 1: beekeeper, syllogism_entity
Entity 2: beekeeper, syllogism_entity

Recall that this description does not imply that the model contains any kind of representation of the word "beekeeper", or indeed any other feature uniquely indicating beekeeperhood. It merely reflects the fact that the model contains representations of entities which can be described as beekeepers, with the details of the representational primitives being inaccessible within the system.[9] Under the program's all-or-nothing template matching this actually means that the attributes and values describing these individuals are a superset of those associated with the word "beekeeper" (although in general the match need not be that close).

Once some beekeepers have been added to the model, the interpretation of the first premise continues by gathering together the set of artists. Since this is the first mention of artists, this set is also empty. The intersection of this set with the set of beekeepers is then computed, which reveals that as the model stands it is not the case that all the beekeepers are artists – none of them are. As a result, the procedures for making sentences true invoke the procedures for adding properties to models in order to modify the description of each of the beekeepers in order to make them also fit the template associated with the word "artist". If artist and beekeeper had conflicting templates, this attempt would fail and the program would reject the premise as incomprehensible. However, the words used in normal syllogistic problems are taken to be unrelated, so the modification can proceed, giving rise to a model which the program would describe thus:

Entity 1: artist, beekeeper, syllogism_entity
Entity 2: artist, beekeeper, syllogism_entity

Finally, subjects are by default modelled as recognising that A premises are not convertible, which in this case amounts to considering the existence of further artists that are not beekeepers. When these too are created, the model for the first premise is complete:

---

[8] The program also counts the number of times it attempts to make the premises true, and declares that the premises are inconsistent if a suitable model is not constructed by the third repetition.

[9] In fact, as illustrated by the example model in Section 7.2.2.4, this will in fact involve noting that they are each "wearing a silly hat", or in general adding the appropriate identifier to a property list. However, even when such a symbolic approach is adopted, the model manipulation routines available do not allow the higher level processes to make any use of it – that is, they are given no facilities to allow any kind of ordering or negating of properties.

Entity 1: artist, beekeeper, syllogism_entity
Entity 2: artist, beekeeper, syllogism_entity
Entity 3: artist, syllogism_entity
Entity 4: artist, syllogism_entity

It should be emphasised that because the words that can be used to describe entities within models may not be explicitly indicated, the model itself imposes no kind of ordering on them. The model simply indicates the existence of entities which will (or will not) fit particular templates.

Once the model has been modified to capture a situation compatible with the first premise, the second premise – "no chemists are beekeepers" is processed in the same way. As before, the first step is to identify the set of chemists, and since there are again none, the program creates some, giving rise to the following model:

Entity 1: artist, beekeeper, syllogism_entity
Entity 2: artist, beekeeper, syllogism_entity
Entity 3: artist, syllogism_entity
Entity 4: artist, syllogism_entity
Entity 5: chemist, syllogism_entity
Entity 6: chemist, syllogism_entity

It then goes on to find the set of beekeepers (which in this case is Entity 1 and Entity 2) and find its intersection with the (newly created) set of chemists. Since this is empty, nothing more need be done – the model now captures a situation in which the second premise holds.

Finally, as mentioned above, the program checks that both the premises do in fact now hold in the model, in case making the second one true had in some way invalidated the first. Since they do, the reading of the premises is complete, and the program starts trying to find a possible conclusion using the generate-and-test strategy described above.

The program's generation of potential conclusions involves the application of the set of conclusion generation strategies it has been given. This discussion will assume that the only strategy the program has at its disposal is the one for which the actual code appears in Figure 7.1. In the discussion of conclusion generation strategies in Section 7.2.2.3.1, it was glossed as:

IF the second premise has the form "all Bs are Cs"
THEN a possible conclusion can be found by
replacing any B in the first premise by C.

The first thing the program does is to test whether this strategy applies to the premises as they are presented – i.e. to

all beekeepers are artists
no chemists are beekeepers.

However, the second premise is not of the required form ("all Bs are Cs") so the strategy cannot be used. Next, the program reverses the order of the premises, and attempts to match against

no chemists are beekeepers
all beekeepers are artists.

This does match, suggesting that a potential conclusion can be found by replacing all instances of "beekeeper" in "no chemists are beekeepers" by "artists". Doing this generates "no chemists are artists" as a potential conclusion.

Having been suggested, this candidate conclusion is checked for pragmatic acceptability. In the default model for a subject, this involves checking that it does not contain the middle term – the term which is common to both premises, which in this case is "beekeeper". Since it does not, and since the sentence is actually true in the model, the program decides that it has successfully generated a candidate conclusion, and passes it to the testing phase of the overall generate-and-test loop.

Testing the conclusion involves attempting to falsify it by modifying the model in ways suggested by a collection of falsification strategies. In the current example, the subject is assumed to possess only one such strategy – namely that which suggests falsifying **E(x, y)** by trying to make **I(x, y)** hold. Fortunately, this strategy fits the candidate conclusion under consideration, and the program therefore takes the suggested action. It calls the procedure for making sentences true within the model, just as if "some chemists are artists" had been a third premise in the problem. As before, making this sentence true involves identifying the set of artists and finding its intersection with the set of chemists. Since the intersection is empty, the model must be modified in order to make the sentence true. The program does this by making some additional chemists, some of whom are also artists, giving:

Entity 1: artist, beekeeper, syllogism_entity
Entity 2: artist, beekeeper, syllogism_entity
Entity 3: artist, syllogism_entity
Entity 4: artist, syllogism_entity
Entity 5: chemist, syllogism_entity
Entity 6: chemist, syllogism_entity
Entity 7: artist, chemist, syllogism_entity
Entity 8: chemist, syllogism_entity

The program now tests that it has falsified the generated conclusion – i.e. the model reflects a situation in which the conclusion is now false, but the premises are true. This is indeed the case, so the conclusion has been rejected.

Having falsified this possible conclusion, the program restores the model to the way it was before it was modified[10] and returns to the task of trying to generate others by matching its strategies against permutations of the premises. Since the actual premises have now been tried in both orders, it now begins to re-arrange the words within the premises. First it swaps the terms in the second premise, giving :

all beekeepers are artists
no beekeepers are chemists

but, as with the original premises, this does not match the conclusion generation strategy. Then this premise pair is reversed, and the program tries to generate a conclusion from

[10] It would also have been possible to arrange for the program to continue working with the model as it had been left by the process of falsifying the conclusion. Since the theory has nothing to say on this matter, the program resets the model every time it defeats a conclusion because this is a more restricting strategy. Resetting the model corresponds to the effect of the subject simply re-reading the premises, which is something they could do at any time. However, by always doing it, the program loses out on any advantage that subjects could gain by working with more refined models of the situations. Although the adoption of this approach was an arbitrary decision, it has been held constant for all subjects modelled.

no beekeepers are chemists
all beekeepers are artists.

Once again, the conclusion suggesting strategy matches, only this time replacing "beekeepers" by "artists" yields "no artists are chemists" – the converse of the previous suggestion. As before, the program checks that the suggested conclusion is pragmatically acceptable and true in the model, and accepts it as a candidate for falsification. The same falsification strategy is once more applicable, this time suggesting that the model should be modified to make "some artists are chemists" hold. Once again, the sets of artists and chemists in the (re-set) model are evaluated, and found not to have any members in common, and as a result, the program creates a set of additional artists and makes some of them chemists, thus:

Entity 1: artist, beekeeper, syllogism_entity
Entity 2: artist, beekeeper, syllogism_entity
Entity 3: artist, syllogism_entity
Entity 4: artist, syllogism_entity
Entity 5: chemist, syllogism_entity
Entity 6: chemist, syllogism_entity
Entity 9: artist, chemist, syllogism_entity
Entity 10: artist, syllogism_entity

As before, this modification serves to defeat the postulated conclusion, and the program returns once more to searching for another.

The last option open to it is to reverse the constituents of the first premise, giving "all artists are beekeepers". This is not a logically valid step, but this does not matter, since it is not intended to be a logical operation. It is not part of an attempt to **deduce** a conclusion, but simply a ploy which might help trigger the recognition of one.

Having produced this altered first premise, the program goes on to consider combined in both possible positions with both possible orderings of the other premise. However, none of these gives rise to a pragmatically acceptable conclusion, and having tried them, the program has exhausted its available techniques for generating conclusions. As a result, it is obliged to give up trying to solve the problem, and it finally declares that it can find "no valid conclusion".

In fact, this is a very common response to this particular problem. It is also incorrect. Given the interpretation of universal quantification over empty sets normally applied in syllogistic reasoning (see Section 4.2.2), "some artists are not chemists" is a valid conclusion. It is important to realise why the program did not give this conclusion. It was not prevented by any limitation on its modelling ability or its grasp of language – indeed it is perfectly capable of **validating** the conclusion. It did not offer it because it did not possess any strategy that suggested it – in other words, it didn't think of it! This is a concrete illustration of the kind of ability error (as opposed to performance error) that the new theory argues for.

7.2.2.5. Modelling the Experimental Evidence

The conclusions offered by the program clearly show a number of characteristic effects that have been noted in the responses of human subjects. Some of these can be unequivocally be associated with a particular strategy that subjects are postulated to be using. For instance, the fact that the presence of an A premise tends to facilitate drawing conclusions, (which Lee (1983b)

dubbed the *A-Effect*), can be directly tied to the widespread use of a particular of the conclusion generation strategy – namely the very one that is advanced as typical in Section 7.2.2.2.1.

On the other hand, even though the responses also clearly show the figural effect, and indeed the variation on it that puzzled Johnson-Laird and Bara, this cannot be associated with any particular strategy. Instead it emerges from the way that the potential conclusions that subjects consider are formed from the components of the premises, which means they tend to share both sub-components and their roles. For instance, in the example of the **2AE** syllogism, described in detail in the previous section, "no chemists are artists", the possible conclusion which agrees with the figural effect, is considered before its converse "no artists are chemists". As a result, had they both been acceptable to the program, the one which supports the figural effect would have been chosen in preference to the one that runs counter to it. The former can be generated just by considering the premises in the reverse order, whereas the latter only emerges as a candidate after the internal structure of the premises has been altered. The closest one could get to tying down the origins of these effects, therefore, would be to say that they arise as a natural consequence of the fact that subjects consider the premises the way they are presented before they consider re-arrangements of them.

The way the simulator is written, all strategy selection decisions (for premise permutation, conclusion generation or model modification) are as simple as possible. The Prolog backtracking mechanism is used to ensure that the spaces of possible conclusion generations and falsifications are exhaustively searched, which has the effect that possible strategies are tried in a fixed order, with applicable strategies being pursued immediately. This is adequate to ensure that any appropriate strategy will always be tried, but is unable to provide any flexibility in the order in which conflicting strategies are considered. However, attempting to model individual subjects, and particularly their behaviour in the symmetrical figures (i.e. 3 and 4), suggests that this is too simplistic. It is necessary to assume that they employ some kind of mechanism for explicitly deciding which premise permutation to consider next, or which strategy's suggestion to evaluate first.[11] This decision could well be influenced by the evaluation of the options against criteria outside the strict limits of the content of the particular syllogism. For instance, the decision may be made on the basis of the apparent relative effectiveness of the various procedures at that particular moment, or the balance between the perceived effort involved and the subject's instantaneous level of enthusiasm for the task.

In most cases the strategies available to the subject will support only a single form of conclusion to any premise pair. In these circumstances, altering the order in which strategies are considered will only affect the amount of work required to find that conclusion, and this is not a feature of the program which is intended to be significant. In some cases, particularly in the symmetrical figures (i.e. 3 and 4), the subject could support more than one conclusion, and the order in which options are pursued determines which is found first. As the results presented in the

[11] There is, of course, no necessity for these decisions to be taken consciously – that is, although they may be very complex computations, they could take place outside the flexible autonomous system responsible for consciousness.

next section show, the program's rigid approach to the order in which it considers strategies is a noticeable limitation on its ability to mimic the precise details of a subject's behaviour for such items.

7.3. Modelling Specific Subjects

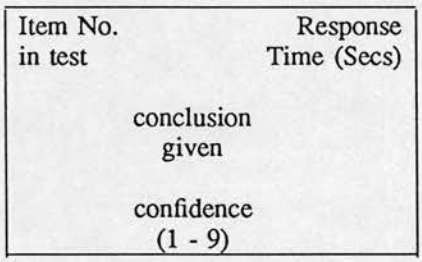
Having described the theory and the machinery, there is enough context in place to consider some results. The program described above has been used to illustrate the consequences of various combinations of strategies, and some have been found which give rise to (close approximations to) the observed performances of a number of subjects. Since it is difficult to envisage all the effects of any given change, this is a time consuming process, requiring many iterations each involving the computer solution of a complete batch of problems. The analyses of the response patterns for three subjects will be presented: two from subjects doing the syllogism task for the first time and one set from a subject performing the experiment for the fifth time. For the first subject, the steps followed in formulating the model will be outlined, whereas for the later subjects, only the model produced will be presented.

7.3.1. Subject A

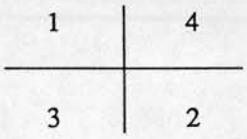
Table 7.1 contains the full results given by Subject A (for Key, see Table 7.0). One response

---

Each entry indicates the following information:



Johnson-Laird's figures are arranged as:



Premises and conclusions are encoded according to the following scheme:

A(x, y) encodes "all x are y". **E(x, y)** encodes "no x are y".
I(x, y) encodes "some x are y". **O(x, y)** encodes "some x are not y".

NVC denotes the subject stating that there is No Valid Conclusion.
- is equivalent to *nvc* in the compressed forms.
???? shows the garbled response by Subject A.

Table 7.0: Key for Subject Result Table Entries.

A(a, b)	**I(a, b)**	**E(a, b)**	**O(a, b)**		**A(b, a)**	**I(b, a)**	**E(b, a)**	**O(b, a)**
30 24 A(a, c) 9	50 20 I(a, c) 9	25 22 E(a, c) 7	27 13 O(a, c) 9	**A(b, c)**	40 13 I(c, a) 7	44 27 I(a, c) 7	55 71 E(c, a) 7	6 63 O(c, a) 9
28 26 NVC 7	54 51 NVC 6	43 19 NVC 7	45 65 NVC 3	**I(b, c)**	8 7 I(c, a) 8	3 13 NVC 9	48 37 O(c, a) 5	16 84 NVC 8
19 18 E(a, c) 8	34 65 NVC 4	13 49 NVC 9	37 79 NVC 3	**E(b, c)**	52 14 NVC 9	39 8 NVC 9	31 32 NVC 7	2 21 NVC 9
64 27 O(a, c) 4	14 33 NVC 5	20 153 NVC 4	49 64 NVC 7	**O(b, c)**	56 13 O(a, c) 5	21 18 NVC 7	33 88 NVC 3	10 13 NVC 8
A(a, b)	**I(a, b)**	**E(a, b)**	**O(a, b)**		**A(b, a)**	**I(b, a)**	**E(b, a)**	**O(b, a)**
26 39 NVC 8	7 30 NVC 7	58 37 E(c, a) 7	18 35 NVC 7	**A(c, b)**	59 A(c, a) 8	32 12 NVC 5	63 52 E(c, a) 4	9 19 O(c, a) 7
1 26 NVC 9	38 15 NVC 9	47 46 O(c, a) 7	29 57 NVC 4	**I(c, b)**	17 16 I(c, a) 8	22 30 NVC 8	4 51 ???? 7	23 15 NVC 7
11 60 E(c, a) 9	60 41 O(a, c) 6	5 21 NVC 9	61 20 NVC 7	**E(c, b)**	42 110 NVC 8	15 51 NVC 7	62 74 NVC 5	51 17 NVC 7
41 60 O(c, a) 6	36 23 NVC 6	12 32 NVC 7	24 5 NVC 7	**O(c, b)**	35 O(c, a) 7	53 9 NVC 8	46 40 NVC 7	57 13 NVC 8

Table 7.1: Full results for Subject A. For Key, see Table 7.0

(**2EI** – the fourth test item the subject encountered) is considered as an "aberration", since the subject's actual response could not be unambiguously categorised into one of the requested forms: Henceforth, it is treated as a response of "no valid conclusion" (*nvc*). These results are repeated, in a more compact format, in Table 7.2(i).

The first feature to notice about the responses is that the 28 premise pairs containing an **A** premise provide 20 of the 23 conclusions offered, including all of the invalid ones. This suggests

that an **A** premise might well facilitate conclusion offering – the *A-Effect* of Lee (1983b). In fact, the conclusion generation rule used as an example in the program description in the previous section (substituting predicates from an "all" premise into another "congruent" premise) fits well with this. When combined with very straightforward model making and some means of rearranging the premises to try all possible orderings, this gives 22 of the 28 responses correctly, and another three correct except for the predicate order (see Table 7.2(ii)). Adding enough skill to try "some X are Y" when "all X are Y" has been found inappropriate corrects **4AA**, and leaves three types of response to be explained: conclusions offered for problems not involving **A** premises, the **E**s at **4AE** and **2AE** and the *nvc* at **3OA**.

When seeking a strategy for getting conclusions from premise pairs that do not have an **A** premise, there is a significant problem. All the conclusions offered are for premise pairs containing an **I** and an **E**. However, although all eight such premise pairs permit a conclusion, the subject only offers one for three of them. It would undoubtedly be possible to find some combination of rules that generate all and only these three solutions. However, the subject's solution-seeking arsenal currently includes the strategy of considering all possible permutations of premises and predicates. If nothing more is done, the subject would be expected to rearrange the remaining five premise pairs until they matched the rules, and thus produce a valid (undefeatable) conclusion. Thus the model of the subject would have to be revised to limit the ability to re-order the premises, and the "A-effect" rule made correspondingly more complicated in order to compensate.

However, there is another feature of the subject's responses to the eight combinations of **E** and **I** premises: as the subject dealt with these eight (one at a time, scattered at random throughout the total batch), a conclusion was only offered for the **last three**. This strongly suggests that, rather than possessing a specific, asymmetrical strategy that works only for some such combinations of premises, the subject uses a general strategy that will work for any of them, but only **acquires** it two thirds of the way through the experiment, as a result of the experience of attempting the first 40 or more examples. If this is idea is accepted, then there is no need to revise the treatment of premise pairs containing "all" since the new rule is itself symmetric.

Now consider the remaining inaccuracies of the computer model. The invalid **E** conclusions being offered for the **2AE** and **4AE** syllogisms should be no surprise, since the subject as modelled so far makes no attempt (or has no idea how) to disprove tentative conclusions. Simply adding some ability to test **E** conclusions, by trying to make a falsifying group of individuals, will easily correct them both. Unfortunately, it will also defeat the desired (observed) but invalid **E** conclusions for **1EA** and **4EA**. However, studying the data of Table 7.1 reveals that the subject spent 110 seconds studying the **2AE** problem before deciding that there was no conclusion. It therefore seems reasonable to suggest that the result of this considerable effort was the recognition of the method of attempting to falsify **E** conclusions. If such a strategy is introduced at this point, the model no longer gives the unwanted responses to the **4AE** and **2AE** forms, but it does still produce the one given by the subject for the **1EA** remains. Unfortunately adopting such a strategy also allows the model to defeat the **E** conclusion which the subject in fact offered in response to the **4EA**. This discrepancy is not easy to fix and remains a misfit between the model and the

	A(a, b)	I(a, b)	E(a, b)	O(a, b)		A(b, a)	I(b, a)	E(b, a)	O(b, a)
A(b, c)	A(a, c)	I(a, c)	E(a, c)	O(a, c)	**A(b, c)**	I(c, a)	I(a, c)	E(c, a)	O(c, a)
I(b, c)	—	—	—	—	**I(b, c)**	I(c, a)	—	O(c, a)	—
E(b, c)	E(a, c)	—	—	—	**E(b, c)**	—	—	—	—
O(b, c)	O(a, c)	—	—	—	**O(b, c)**	O(a, c)	—	—	—
	A(a, b)	**I(a, b)**	**E(a, b)**	**O(a, b)**		**A(b, a)**	**I(b, a)**	**E(b, a)**	**O(b, a)**
A(c, b)	—	—	E(c, a)	—	**A(c, b)**	A(c, a)	—	E(c, a)	O(c, a)
I(c, b)	—	—	O(c, a)	—	**I(c, b)**	I(c, a)	—	—	—
E(c, b)	E(c, a)	O(a, c)	—	—	**E(c, b)**	—	—	—	—
O(c, b)	O(c, a)	—	—	—	**O(c, b)**	O(c, a)	—	—	—

7.2(i) Subject As actual responses (compressed form of Table 7.1).

	A(a, b)	I(a, b)	E(a, b)	O(a, b)		A(b, a)	I(b, a)	E(b, a)	O(b, a)
A(b, c)	A(a, c)	I(a, c)	E(a, c)	O(a, c)	**A(b, c)**	I(c, a)	I(c, a)	E(c, a)	O(c, a)
I(b, c)	—	—	—	—	**I(b, c)**	I(a, c)	—	—	—
E(b, c)	E(a, c)	—	—	—	**E(b, c)**	E(a, c)	—	—	—
O(b, c)	O(a, c)	—	—	—	**O(b, c)**	O(a, c)	—	—	—
	A(a, b)	**I(a, b)**	**E(a, b)**	**O(a, b)**		**A(b, a)**	**I(b, a)**	**E(b, a)**	**O(b, a)**
A(c, b)	—	—	E(c, a)	O(a, c)	**A(c, b)**	A(c, a)	—	E(c, a)	O(c, a)
I(c, b)	—	—	—	—	**I(c, b)**	I(c, a)	—	—	—
E(c, b)	E(c, a)	—	—	—	**E(c, b)**	E(c, a)	—	—	—
O(c, b)	O(c, a)	—	—	—	**O(c, b)**	O(c, a)	—	—	—

7.2(ii) Solutions for A-effect strategy.
(with checking for truth in initial model)

	A(a, b)	I(a, b)	E(a, b)	O(a, b)		A(b, a)	I(b, a)	E(b, a)	O(b, a)
A(b, c)	A(a, c)	I(a, c)	E(a, c)	O(a, c)	**A(b, c)**	I(c, a)	I(c, a)	—	O(c, a)
I(b, c)	—	—	—	—	**I(b, c)**	I(c, a)	—	O(c, a)	—
E(b, c)	E(a, c)	—	—	—	**E(b, c)**	—	—	—	—
O(b, c)	O(a, c)	—	—	—	**O(b, c)**	O(a, c)	—	—	—
	A(a, b)	**I(a, b)**	**E(a, b)**	**O(a, b)**		**A(b, a)**	**I(b, a)**	**E(b, a)**	**O(b, a)**
A(c, b)	—	—	E(a, c)	—	**A(c, b)**	A(c, a)	—	E(c, a)	O(c, a)
I(c, b)	—	—	O(c, a)	—	**I(c, b)**	I(c, a)	—	—	—
E(c, b)	E(a, c)	O(a, c)	—	—	**E(c, b)**	—	—	—	—
O(c, b)	O(c, a)	—	—	—	**O(c, b)**	O(c, a)	—	—	—

7.2(iii) Solutions for Subject A (as modelled by program).

Table 7.2: Conclusion tables for Subject A. For Key, see Table 7.0

---

subject – an execution error.

The **O** conclusion that the model (but not the subject) offers for the **3OA** is actually valid, so it cannot be removed by sharpening up the subject's logical ability (this same argument also indicates why the unwanted conclusions to combinations of **I** and **E** premises had to be dealt with by limiting the generation of candidate solutions). In fact, quite the opposite, the aim is to deliberately introduce ineffective reasoning, preferably in a carefully controlled manner. The method chosen for doing this is to postulate that the "all" substitution cannot occur into predicates

of the form "some X are not Y". This is in effect suggesting that, initially at least, the subject took the premises to be "some X are not-Y" and "all Z are Y", and thus did not fully appreciate the connection between "not-Y" and "Y".

If the suggestions put forward above are followed, the set of rules one arrives at produce the solutions presented in Table 7.2(iii). These differ from the actual responses in two ways: The subject erroneously offered an **E** conclusions to the **4EA** problem while the program does not, and the model reverses the premise order in three conclusions (**4IA**, **3EA**, **3AE**). The latter can be fixed by modifying the generator to ensure that the permutations of premise words are considered in a different order, but this has the effect of also reversing the order of the terms in other forms. This shows that this is a very specific type of discrepancy – the subject is applying a set of rules in varying order, which the program cannot mimic – which was mentioned in Section 7.2.2.3.

In the light of this model, how can the subject's performance be described? To start with the only strategy the subject had for formulating a possible conclusion was to substitute one predicate for another when prompted to do so by an **A** premise – a modest initial ability in line with the unfamiliarity of the task. Similarly, since disproving deductions is also unfamiliar, the subject has no effective way of doing this either. After twenty problems, 153 seconds spent on the 1EO problem leads to the realisation of the connection between "Y" and "not-Y". Another twenty problems later, and 110 seconds playing with the **2AE** leads to a recognition of what to do to falsify an **E** conclusion, although it is not enough to uncover the possibility of trying an **O** conclusion (which would have been valid) instead. Finally, soon after that the subject discovers that an **E** premise and an **A** premise can combine to give an **O** conclusion. At the end of the hour, and as a result of grappling with 64 challenging problems, the subject has improved in this unfamiliar task.

7.3.2. Subject B

The results from Subject A give the impression of steady, monotonic progress towards expertise, being reflected in steadily improving results. The path to competence followed by Subject B, in contrast, exhibits apparent regressions in ability. The details of her results are shown in Table 7.3. This time the description will simply focus on the finished model, although of course there are simple rules underlying each ascription of a "belief" or "tendency".

When a computer model of this subject was being constructed, a number of features of the subjects performance influenced the process.

(1) This subject gives three conclusions including the middle term, and rather than just dismissing these as performance errors, the model attempts to incorporate them. These responses all occur during the early stages of the experiment, which suggests that the subject decides, or learns, during the experiment that they are unsuitable (recall that the garbled response made by Subject A was on his 4th test item).

(2) During the first 24 responses the subject only offered **I** ("some") conclusions. This suggested that the subject was initially treating the task as one of determining whether some entities with one property also had another – i.e. only considering whether or not "some A are C" was suitable. This is captured by a model that starts out generating only **I**

A(a, b)	**I(a, b)**	**E(a, b)**	**O(a, b)**		**A(b, a)**	**I(b, a)**	**E(b, a)**	**O(b, a)**
44 31 A(a, c) 9	45 11 I(a, c) 9	60 56 NVC 9	43 90 NVC 6	**A(b, c)**	61 28 I(a, c) 7	4 21 I(c, a) 9	1 142 NVC 7	24 27 O(c, a) 9
21 24 NVC 8	42 30 NVC 8	27 29 O(c, a) 8	33 19 NVC 8	**I(b, c)**	2 41 I(c, a) 9	59 11 NVC 9	29 13 O(c, a) 9	34 22 NVC 7
54 14 E(a, c) 9	48 53 NVC 7	13 34 NVC 7	30 32 O(a, c) 9	**E(b, c)**	40 54 NVC 7	18 32 O(a, c) 9	10 40 NVC 6	28 111 NVC 5
58 27 NVC 8	12 22 NVC 6	31 79 NVC 4	11 49 NVC 6	**O(b, c)**	47 51 O(a, c) 8	35 30 NVC 8	26 34 NVC 7	20 29 NVC 7
A(a, b)	**I(a, b)**	**E(a, b)**	**O(a, b)**		**A(b, a)**	**I(b, a)**	**E(b, a)**	**O(b, a)**
39 44 NVC 7	49 35 NVC 7	25 26 E(c, a) 9	56 35 NVC 8	**A(c, b)**	22 24 I(c, a) 9	52 22 NVC 7	36 46 E(c, a) 9	51 54 NVC 7
9 58 I(b, a) 6	6 44 NVC 4	63 100 O(a, c) 5	5 45 I(b, a) 1	**I(c, b)**	19 14 I(c, a) 9	53 19 NVC 9	16 23 O(c, a) 9	23 22 NVC 9
46 32 I(c, a) 9	50 28 NVC 8	38 31 NVC 7	14 47 NVC 7	**E(c, b)**	17 30 NVC 8	37 42 NVC 7	32 42 NVC 7	7 46 NVC 8
3 182 I(b, c) 4	55 22 NVC 9	64 122 O(c, a) 5	8 29 NVC 8	**O(c, b)**	15 16 I(c, a) 7	57 14 NVC 8	62 48 NVC 7	41 56 NVC 7

Table 7.3: Full responses for Subject B.
For Key, see Table 7.0

conclusions, and only later also considering other possible forms.

(3) The subjects handling of **EI** and **IE** syllogisms shows a striking temporal pattern: Initially she gives valid **O** conclusions, then she goes through a phase of saying there is no conclusion, and then, right at the end of the experiment, she offers a converted (and thus invalid) **O** conclusion. She produces similar behaviour with her responses to **EO** and **OE** syllogisms, which have no valid conclusions. Here also she initially gives **O** conclusions, then says there

is no conclusion, then resumes giving O conclusions. What is most striking, however, is the fact that the O conclusions to the two problem forms stop and re-start together! This strongly suggests that there is a common mechanism responsible for generating them which for some reason becomes inoperative during the middle portion of the experiment.

Working from these observations, a model was constructed which suggests the following picture of the subject's behaviour during the experiment. Initially, she believes that the problem is a matter of determining whether or not a conclusion of the form "some a are c" is applicable. However, at the third problem she spends a long time (180 seconds) on the 3AO syllogism, and comes to the conclusion that giving a response involving the middle term is better than nothing. As a result, she adopts a strategy of considering such a conclusion when a "real" one is not appropriate. However, she will not offer a conclusion that is just one of the premises or its converse, and this is soon refined to capture the belief that "some a are b" is effectively just re-stating "some a are not b". The idea of considering middle-term conclusions is finally abandoned somewhere between the 9th and the 16th, as encountering more problems gives her a better feel for the task. In the model this necessarily happens abruptly, though it may well have been a gradual shift of attitude.

Although the subject starts off considering only **I** conclusions, at the 14th item she begins to recognise that other conclusions might be applicable. In particular, she adopts a strategy for generating possible conclusions that is can be captured by

> IF the second premise has the form no Bs are Cs
> THEN a possible conclusion can be found by
> replacing any B in the first premise by not C.

This strategy, which is closely related to that capturing Lee's A-effect (described above), allows her to generate the correct conclusions to **IE** and **EI** syllogisms – i.e. she starts to correctly solve three-model problems. Then, ten items later, she supplements this by adopting the A-effect strategy itself (i.e. when one premise is an **A** form, replace every occurrence of its subject in the other premise by its object).

At item 28, she spends 111 seconds on the **1OE** problem, and begins to associate **O** and **I** premises more closely, and to use this same strategy that she acquired at item 14 for generating conclusions to **OE** and **EO** problems. However, shortly afterwards, 79 seconds spent on the **1EO** problem allows her to see a procedure for defeating particular (i.e. "some...") conclusions. Specifically, she realises that "Some A are B" can be defeated by modifying the model in order to make the statement "No A are B" true. As a result, she realises that there is something wrong with the way she has been handling **OE** problems. Unfortunately, she does not recognise that the problem lies with her linking of **O** and **I** conclusions. Instead, she ascribes it to the conclusion suggesting heuristic, which she decides to stop using. This corrects the problem that she has just discovered – giving invalid O conclusions to **OE** premise pairs – but has the unfortunate effect that she stops giving **O** conclusions altogether! This continues until the penultimate item in the experiment, when having spent 100 seconds studying the **3EI** problem, she finally decides that the discredited strategy did have some merit. As a result, she re-adopts it, and gives an **O** conclusion. However, this reinstated strategy immediately leads her to consider an erroneous conclusion on the

next (and final) item (the **3EO**), and since she has still not sorted out the real source of her problems – the relation between **I** and **O** premises – the 122 seconds she spends on it don't enable her to defeat it.

The computer model that behaves in line with this description produces the responses given by the subject with three exceptions, where in each case the problem is with the term ordering. Two of these (**4AA** and **3OA**) are (logically convertible) **I** responses, and the error appears, in the terms of the model, because the possible strategies are applied in a different order. The third discrepancy, the **3EI** syllogism just mentioned, also has symmetrical premises, but the **O** conclusion is not logically convertible. The conclusion the subject gives is invalid, whereas the model predicts that she should have produced the correct response. This discrepancy is plausibly accounted for by suggesting that it is an execution error, particularly in the light of the fact that she correctly solved the other three (logically equivalent) **EO** problems. Moreover, in this case she spent over 100 seconds on the item, quite possibly concentrating not directly on its conclusion, but on its relevance to other problems in terms of solution strategy re-acceptance.

In contrast to the Subject A, some of the learning suggested by the model of Subject B is regressive – it leads to poorer performance. However, this should not be surprising, since any learning occurring during the experiment does so in the absence of any external feedback. Further, the subject has had no opportunity to grasp or explore the logical structure of the task domain. This means that when the subject stumbles across either an inconsistency (e.g. manages to defeat a conclusion generated by a previously trusted method) or a promising new method (e.g applying a generation strategy that works for **I** premises to **O** premises as well), no means are available to evaluate any resulting action. As a result, it is not surprising that wrong decisions (e.g. rejecting a strategy, rather than just restricting its application) are taken.

7.3.3. Subject C

The models proposed for subjects A and B both had to postulate significant changes in strategy in order to describe the observed behaviour. This is in line with the proposed theory, which does indeed predict that learning will occur. However, these changes are predicted to be common because the subjects are new to the task, and each problem encountered contributes significantly to their experience of the field. This means that by the time a subject has performed the experiment many times, their behaviour should have settled down and the strategies employed should have become stable. The most extreme case of this was Subject C, who gave virtually identical responses on the fourth and fifth times he sat the experiment. This strongly suggests that there is no learning going on, and thus it should be possible to model this particular subject without any kind of strategy changes.

Table 7.4 shows the response pattern that Subject C gave the fifth time he tackled a batch of syllogisms. These results were identical to those he gave the fourth time he took part in the experiment, with two exceptions. The first is that in the fourth sitting, he said that no valid conclusion followed from the **3OE** syllogism, while on every other sitting he offered **O(a, c)**. This can be written off as an execution error on the fourth sitting, which is especially likely in view of the fact that it was the first item tackled on that occasion. The second difference is that the **3EA**

A(a, b)	**I(a, b)**	**E(a, b)**	**O(a, b)**		**A(b, a)**	**I(b, a)**	**E(b, a)**	**O(b, a)**
25 10 A(a, c) 8	38 26 I(a, c) 8	26 16 NVC 7	19 40 I(a, c) 7	**A(b, c)**	34 29 A(a, c) 7	12 21 I(a, c) 7	21 15 NVC 7	57 25 I(a, c) 7
43 36 O(c, a) 5	62 10 NVC 7	47 28 O(c, a) 7	48 14 NVC 7	**I(b, c)**	51 11 I(c, a) 7	55 7 NVC 7	46 13 O(c, a) 7	61 10 NVC 7
9 13 E(a, c) 8	24 18 O(a, c) 7	2 12 NVC 7	54 32 O(a, c) 7	**E(b, c)**	29 10 NVC 7	3 21 O(a, c) 7	52 13 NVC 7	44 22 O(a, c) 7
59 10 O(c, a) 5	14 13 NVC 7	58 12 O(c, a) 7	31 10 NVC 7	**O(b, c)**	17 35 I(c, a) 7	49 29 NVC 7	13 60 O(c, a) 7	1 24 NVC 7
A(a, b)	**I(a, b)**	**E(a, b)**	**O(a, b)**		**A(b, a)**	**I(b, a)**	**E(b, a)**	**O(b, a)**
36 24 NVC 7	32 47 O(a, c) 6	39 12 E(c, a) 8	11 158 O(a, c) 7	**A(c, b)**	27 20 A(c, a)	28 74 O(a, c)	41 11 E(c, a)	40 49 O(a, c)
33 29 O(c, a) 6	64 5 NVC 7	30 21 O(c, a) 7	42 31 NVC 7	**I(c, b)**	7 6 I(c, a) 7	8 NVC 7	60 9 O(c, a) 7	6 13 NVC 7
5 18 E(a, c) 7	45 17 O(a, c) 7	53 15 NVC 7	22 29 O(a, c) 7	**E(c, b)**	16 17 NVC 7	4 44 O(a, c) 7	23 49 NVC 7	18 44 O(a, c) 7
8 32 O(c, a) 7	15 16 NVC 7	50 14 O(c, a) 7	20 7 NVC 7	**O(c, b)**	10 11 I(c, a) 7	37 27 NVC 7	63 12 O(c, a) 7	35 8 NVC 7

Table 7.4: Full results for Subject C
For Key, see Table 7.0

syllogism elicited **E(c, a)** the fifth time and **E(a, c)** the fourth time, and in this seems to reflect a minor inconsistency in the subject's approach, since both and only these responses were observed in the first three sittings. Once again, this is a reversible conclusion from symmetrical premises.

The most striking thing about this subject's responses is the number of **O** conclusions, both valid and invalid. Further examination reveals that in every case, **I** and **O** premises are treated identically. This might lead to the suggestion that the subject considers the two premises

completely synonymous, but there is evidence that this is not the case. The subject offers both **I** and **O** conclusions, and the patterns in the responses show that he is making a methodical choice between them, and it is one that is not simply related to any surface features of the premises.

Starting from this observation, a model was built up for the subject which manages to account for this peculiar behaviour using an unchanging set of rules. It allows him to be seen as having a normal grasp of the meanings of all the sentences involved. That is, he has the ability both to verify any syllogistic sentence in line with accepted logic and to interpret any such sentence by building a model in which it holds. He also possesses a range of sound strategies for generating and testing any form of conclusion, which is what might be expected of someone who has spent several hours doing just these things. Yet in spite of this, it is still able to produce the erroneous conclusions offered by the subject. They all arise from the fact that, when testing conclusions, he will not modify his mental model if doing so will falsify one of the premises.

This defective falsification strategy can be seen as an example of the artificial intelligence concept of *hill climbing* (see Minsky, 1963. P410-411). The term is used to invoke an analogy with attempting to get to the peak of a mountain by always going upwards. It describes a process that tries to achieve the best state of affairs by continually attempting to adjust things in order to effect the maximum improvement. However, things that behave in this way are vulnerable to a very specific but widespread problem concerning local maxima. Scaling a mountain by always going uphill is very likely to result in climbing one of its foot-hills, since there are often places where one must go down – across a depression – to get to the real mountain. Similarly, in seeking an optimum solution there are often cases where the only way to get to a position to make things better requires first making things worse.

Subject C's solution strategy can be readily recognised as an example of hill climbing, and its inadequacy as the result of local maxima. He is seeking to falsify a syllogistic conclusion by producing counter-examples, where both premises must hold. Any step which violates one of them is clearly retrogressive – "downhill". However, in certain cases it is sensible because the premise can then be made true in some other way that leaves the conclusion false. By adopting a strategy of never falsifying a premise, Subject C prevents himself from ever defeating some conclusions.

It is also possible to see why the subject should have adopted such a strategy. When solving many syllogisms subjects will obviously try to minimise the effort involved. Obviously, they will attempt to avoid useless activity, as has already been suggested in the context of conclusion generation. However, in the absence of feedback on the accuracy of their performance, this will amount to abandoning strategies because they **appear** fruitless. The negative aspect of a step that falsifies something that must be true would certainly provide grounds for discrediting a strategy that involves it, and would be very apparent. In contrast, appreciation of the benefit of such a move would require that the subject be able to make the violated premise true again in an alternative way – one that would not also restore the conclusion. However, Waddington's experiment (described in Section 4.2.2) suggests that this is unlikely. In the absence of any effect of lexical material, she found that subjects have a strong tendency to associate quantified sentences with one particular set relationship. Thus it is quite likely that the subject would be unable to discover the usefulness of

this strategy – and thus would reject it – because he lacked the specific ability to explicitly recognise the possibility of interpreting a "meaningless" sentence in more than one way.

Notice that, in line with its development in the absence of feedback, the subject's stable response pattern is not particularly accurate – 42 valid responses in the batch of 64. Nor is it obviously simple – it includes conclusions of all four forms, and some that are against the figural bias. Nonetheless, it can be produced very accurately from a stable set of simple rules. Moreover, the subject's ideas of the meanings of the quantifiers are sound, and the models he builds are in line with common interpretations of the premises.

7.3.4. Summary

All three models just presented achieve extremely accurate matches to the actual responses of the subjects. By far the most common discrepancy relates to conclusions expressing symmetrical relationships (**I** and **E**) in the absence of figural bias. In these cases, the simulator would have given the right conclusion had it not arrived at another one first. The difference between the subject's response and that given by the program can be ascribed to the order in which the permutations of premises were tried. Simply changing the order in which they are tried disrupts other, currently correct, responses. This suggests that the problem arises because the strategies available to the subject could lead to more than one conclusion, and the order in which they are considered decisive. Unfortunately, as pointed out above, the simulator is not designed to cope well with choices. Strategies are exhaustively considered in a fixed order and it is completely algorithmic in its permutation sequencing. As a result the current simulator is unable to capture these effects, and although it could be re-written to do so, the wisdom of doing so is discussed in Section 7.6.

The distribution of the strategy ordering effects is suggestive of the sort of factors involved. The fact that they are confined to the two symmetrical figures (i.e. 3 & 4) indicates that the arrangement of terms might be significant. Any strategies for selecting terms for a conclusion can clearly be applied to the two premises in either order, and the decision would determine the term order in the conclusion. In the asymmetrical figures, the order of consideration would be irrelevant, since one direction of application would result in a pragmatically unacceptable conclusion involving the middle term. But in the symmetrical figures, the two possibilities would be equally effective at generating acceptable conclusions, and thus the ordering would be significant. Moreover, precisely where it is important, there is nothing in the surface structure of the premises to direct the choice, and only such features are relevant[12] . Thus the most common inaccuracy of the models can be seen as arising at points where the subject is predicted to be influenced by factors beyond its range.

If the effects of strategy ordering are recognised as a specific shortcoming, the closeness of fit of the models is excellent. There are only 2 unexplained discrepancies, or less than 1 per batch of 64 syllogisms, and each of these can be identified with a factor that might have induced it.

[12] Newell (1981, P700) cryptically points out that only superficial features can guide the initial strategy selection. Attempting to use the result of analysis leads to an immediate regress when one comes to deploy analysis strategies to select which analysis strategy is appropriate.

Unfortunately, since the proposed theory does not postulate any concrete limits on the number of strategies or learning events, it is unquestionably able to "explain" any conceivable pattern of responses. Moreover, the program was written at an early stage of development of the theory, at which time it was not obvious what features of solving behaviour would vary between subjects. As a result, it includes the ability to vary the approach taken to almost every decision, and as a result, there are potentially many ways of achieving any desired pattern. Fortunately, however, it transpires that only a few of these parameters have actually had to be varied in order to allow a set of strategies and strategy changes to be found to model the responses of actual subjects. These parameters that actually had to be varied are:

(1) The range of conclusion generation strategies used by the subject. As predicted by the theory, most of the flexibility of the system comes from here. Those used to model the subjects presented above were:

The A Effect: When one premise links a predicate to the middle term by a universal quantifier ("all"), substitute the occurrence of the middle term in the other premise by that predicate. This is the strategy implemented by the code in Figure 7.1, and is named after the effect observed by Lee (1983b), to which it gives rise. It forms the basis for all the subject models presented.

Weakening: When the subject considers an **A** conclusion and manages to falsify it, this strategy suggests also considering an **I** conclusions. That is, if ever the program manages to defeat a conclusion of the form "All whatsits are doobries", the "weakening" strategy will lead it to consider "Some whatsits are doobries" instead. It would also be possible to postulate a similar weakening of **E** conclusions to **O** conclusions, but it has not been found necessary to use it.

The E Effect: This is in a sense the complement of the **A** effect. When an **E** premise links the middle term to a predicate, try replacing the middle term in the other premise with the negation of the predicate. Thus if the premise says that "No doobries are whatsits", try replacing any occurrence of doobrie in the other premise with "not whatsits". This is the strategy that is postulated to generate the conclusions offered for **IE** and **EI** premise pairs, for which it gives a (valid) **O** conclusion. However, it can also give rise to other sentences which are not syllogistic forms, for instance because they include double negation ("some of the whatsits are not not doobries"). Once again, it is reasonable to suggest that subjects may vary in what they do under such circumstances (i.e. this is another potential degree of freedom). However, to date all have been modelled as recognising and normalising double negatives in the object position of the proposed conclusion, and simply rejecting any negative in the subject position ("some whatsits are not not doobries" is recognised as "some whatsits are doobries", while "some not whatsits are doobries" is simply rejected as a non-sentence.

I Conclusions: Simply consider an "I" conclusion linking two predicates. This strategy is suggested as the manifestation one particular subjects incomplete grasp of what the experiment was about.

(2) The range of conclusion falsification strategies used by the subject. More specifically, whether or not the subject knows how to (or knows it is worthwhile or appropriate to) falsify premises by making the contradictory sentence true.

(3) The subject's precise grasp of the pragmatic (experiment-related) constraints on an acceptable conclusion. In particular, whether subjects know to reject all conclusions containing the middle term.

(4) Whether "not X" is treated as atomic or not by substitution rules – that is, whether the subject would consider substituting for "doobries" in a premise of the form "some of the thingies are not doobries".

(5) Whether or not the subject will consider trying to re-satisfy a premise that has been defeated along with a potential conclusion.

In essence these results are in line with the theory being proposed. It has proved possible to account for most of the detail of the subjects' responses in terms of the presence or absence of each of a small number of task-specific strategies. Admittedly, in order to achieve the very precise fits that are necessary to avoid appealing to any kind of execution error, it has been necessary to vary certain other features to accommodate particular subjects. However, these modifications are confined to details of the procedures which are specific to syllogism solving. Every subject is still modelled as having a good grasp of what is required in syllogistic reasoning and a perfect knowledge of the quantifiers this involves. Moreover, they can reliably entertain a single unambiguous representation of a situation for as long as they need to work on it. Nevertheless, however plausible the resulting stories may be, the fact that it was necessary to adjust features of the subject models on an ad hoc basis does represent a shortcoming of the account. The obvious next step is to attempt to characterise and delimit the range of such adjustments that are needed, primarily by constructing computational models for a large number of subjects. However, there are difficulties with this approach, a matter which is take up again in Section 7.4

Finally, although no attempt has been made to capture the learning processes that give rise to changes in strategy, the fact that they often occur when the subject spends a long time studying one item has been stressed. This does not imply that the process of finding a way to tackle a problem is a protracted one that in some sense holds the subject up. Everyday experience suggests that finding a solution to a problem can be very fast when the problem is approached in the right way. In terms of syllogism solving, the learning process itself might be very fast once the appropriate strategy was recognised. The tendency to learn on problems that are studied intensively may arise because that is when a problem is most likely to be considered in a novel way.

7.4. Specific Features of the New Account

The new theory has many appealing features, stemming largely from the recognition that explicitly drawing a conclusion from specific premises and ensuring that it will always be supportable is an unfamiliar task. As a result, there is no reason to suppose that the subject approaches it with any specific skills or abilities that are relevant (cf. Johnson-Laird's syllogism-oriented conceptual models). Instead, the problem must be solved by combining, at first probably consciously, mechanisms or abilities already acquired for some other purpose. This requirement is

met by the new account, which proposes a mechanism that is based upon simple and very general skills and, moreover, combines them in a very natural way.

The overall strategy that is proposed can be summed up as Starting from the premises, try and find a sentence so that you can't think of a situation where the premises are true and that sentence is false. This is almost a literal interpretation of the instructions that subjects were given at the start of the experiment. In other words, if one imagined the subject trying to work out, from a definition of a syllogism, how to go about getting valid solutions, this is the sort of answer that one might reach. Moreover, the theory makes very clear the way in which the meanings of the quantifiers are employed, in controlling the validation of the sentences. Further, it is obvious in what sense the subjects can be said to understand the problem and the concept of a valid deduction, in that they can be seen to be seeking a solution that is true in every possible situation where the premises are true. In fact, what the subject is actually finding is a solution that is true in every such situation that is conceivable **for them**, with incorrect responses resulting when ability errors lead to a failure to consider a class of situation.

The information structures employed within the new account have less representational power than those proposed by Johnson-Laird, and manage to remain within the limitations proposed in Chapter 1 on the basis of other everyday mental phenomena. However, this does mean that they correspond very closely to a scene or situation, about which there are an infinite number of things that can be said. This makes Johnson-Laird's approach to conclusion generation by ''reading off the model'' quite intractable, since it relies on so structuring the model as to restrict the range of things that are licensed by it (or that one is allowed to say based on it, since these two come apart under his approach).

This intractability is by no means a bad thing, since as pointed out in Section 6.6.2, mental models do not decide what should be said in normal conversation. As a result, there is no reason why syllogistic reasoning should be any different, particularly if the same type of models are underlying the process. The role of the model is not to constrain the range of candidate conclusions, but to provide the means of checking the truth or plausibility of sentences, a much simpler process than deciding what to say. In addition, this approach allows an obvious extension to the closely related problem of validating syllogisms or solving multiple choice problems, by allowing candidate conclusions to be simply read from the card or wherever, instead of being generated. This is in contrast to Johnson-Laird's approach, where there is definite uncertainty concerning how a given conclusion could be validated: would subjects assent to a conclusion that happened to be true in the models they were considering, but could not be read from them?

If the model only has the role of validating, rather than specifying, conclusions, its part can be played by a ''homogeneous'' (as opposed to lexicalised) model. Such models bring with them the advantage of simple interpretation of their contents. Every description of a model entity corresponds to a cluster of data, however represented, that is to be taken as modelling for the model builder some (one) object in the world which is such that it can be ascribed each of the properties mentioned in the description. In line with the restrictions suggested by mental imagery, there is no attempt at representing negative information – an entity in a model, and thus the entity

it models, simply has the properties it has.

The features that make for this simplified interpretation also simplify the extension of the model, since it need only be a "local" phenomenon. Items can be added to the model without reference to any other items, and the alteration of the properties of any particular entity need only maintain consistency within that entity and its immediate environment in the model. This simplicity of extension is central to the everyday role of mental models in the integration of information about the environment. Since premise interpretation is the result of the subject's everyday linguistic abilities, they will be interpreted as ordinary discourse, albeit (since reasoning about groups of strangers on the basis of their hobbies alone is somewhat unusual) about a rather strange situation. As a result there are no problems associated with models relating more than three terms. In addition, it is to be expected that the exact nature of the material in the syllogism will affect both the models built and the results achieved. This is in keeping with observation and experimental results and contrasts with the ambiguity of the Johnson-Laird approach.

Another consequence of viewing syllogistic reasoning as requiring unfamiliar skills is that subjects can be expected to learn from the experience of trying to solve them. Everyday experience shows that there are two main effects of learning – people learn to do things faster and better. Both these effects are discernible in the results of syllogistic reasoning experiments, and the new theory is able to explain why both should arise

The fact that a generator and a filter are together equivalent to a more specialised generator offers a significant insight into the acquisition of expertise. The latter approach clearly involves less work, but requires a specialised procedure or mechanism which, in the case of an unfamiliar task such as syllogism solving, is unlikely to be available. Thus subjects new to syllogistic reasoning will have to adopt the slower approach involving overgeneration. However, one of the most marked effects of practice was that subjects got faster – by as much as a factor of three. Although the mechanisms of learning are as yet unknown, acquiring a skill is surely dependent on having at least a crude method, and in this respect the filtering from over-generalisation would provide a suitable focus for improvement. The observed speed-up can be seen as resulting from subjects refining their procedures for conclusion generation, and learning to consider only candidates that are suitable.[13] Initially subjects might generate sentences of all possible forms and reject those that are not in syllogistic forms, but with practice they learn how to generate only those that are relevant.

The qualitative improvements in subjects' behaviour are also dealt with by the theory. In particular, subjects errors are accounted for in a way that does not undermine their ability to play Monopoly. The subjects have indeed understood the premises properly, are fully aware what a valid conclusion is and have done their best to get one. They often err, but not because of random slips or execution errors, but because of ability errors. Their poor performance is due not to the

[13] This view of the acquisition of expertise is consistent with the observation that chess masters typically examine only a dozen or so continuations from any position. Nevertheless, they can (currently) consistently beat computer programs which explore thousands. Their expertise lies in the ability to consider the **right** dozen. It has also been pointed out that many great scientific breakthroughs are made by people who are new to the field – too new to know that what they are suggesting is not worthy of consideration...

difficulty of the tasks but to their unfamiliarity with it. Since the relevant syllogistic mechanisms are distinct from those employed in everyday thinking, postulating inadequacies of ability does not impugn the subjects logical competence or every day rationality.

Diagnosing ability errors allows a natural explanation of subjects improvement with practice as the symptom of enhancements of the heuristics that they employ. With experience, subjects will develop better (basically more thorough and methodical) strategies both for generating and defeating potential conclusions. Moreover, this mechanism smoothly incorporates the acquisition of expertise, the natural end-point for this process. The model predicts a "limiting case" where the subject has a very accurate set of solution generating strategies in which he has great faith. Thus he would quickly know the correct solution for any syllogism, and would not bother or need to test his suggestions. In other words, he would know the answers by heart. This is not to suggest that the subject has "internalised formal logic" which is then swiftly applied to the problem. Rather, the subject has simply learned the correct answers "by rote", although this was only acquired through a thorough understanding of the problem.

The fact that correct performance is seen as the result of a number of independent skills naturally explains the observations of Simpson and Johnson (1966) and Dickstein (1975). They each found that training aimed at removing one kind of error had no beneficial effects on others. This is to be expected if the effect of the training is seen as the acquisition of another independent skill which is specific to the avoidance of the errors warned against. Similarly, because skills are being acquired, there is no reason to expect that all subjects will have the same set. They will each bring different skills to bear on the problem, and thus quite possibly be affected by different aspects of it. This leads to the prediction of Wilkin's (1928) observation that different people were influenced to different extents by the lexicalisation of the problems.

As a side-effect of recognising learning syllogistic reasoning as the acquisition of relevant heuristics or strategies, the theory is able to identify the precise skills involved in syllogism experiments. Since the process is seen as essentially a generate-and-test cycle, there is obvious opportunity for expertise at each of these stages. The skill most generally overlooked, particularly by Johnson-Laird, is the ability to think up sensible solutions to try. A failure in this area explains subjects failing to find valid solutions. This is exemplified by a subject discussing his failure to solve one particular (valid) syllogism, who said Oh yes, I remember that one. I just stared at it but nothing happened. As suggested above, fabricating potential conclusions from the premises is a likely first blush strategy, which not only explains the figural effect, but also the particular difficulty of the **4AA**, **1EA** and **2AE** problems. In these cases the valid conclusions are particulars, which will be overlooked because they contain the word "some" which simply does not appear in either premise.

This discussion also highlights the very natural way in which the theory can explain Johnson-Laird's figural effect, not to mention the A-effect, the Atmosphere effect and all the other observations of surface patterns in the solutions. The solutions are influenced by the shape of the premises because they are constructed from them. Moreover, and unlike Johnson-Laird's own account, this theory accords very well with his observation that the figure of the premises has an

effect only when a conclusion is being drawn. The effect of figure arises in the very process of formulating the conclusion. The theory even leads one to expect some of the observations that perplex Johnson-Laird and Bara (1984). As mentioned in Section 6.1, they point out that, in symmetric figures, where the conclusion is in the same mood as one of the premises, the term contained in that premise retains its grammatical position. If the conclusion generation process is seen as one of inter-substituting words, the explanation for this is very obvious: The conclusion looks very like one of the premises because it was **made from it**.

Of course, although the theory suggests that the manipulation of verbal representations is involved, it should not be confused with suggestions like the "Mr. Hyde" atmosphere effect – the idea that subjects merely take account of the surface features of the premises with no grasp of the meanings involved. The current theory quite definitely says that subjects are using the meanings of the premises in producing their conclusions. However, it accords well with the original "Dr. Jekyll" atmosphere effect proposed by Woodworth and Sells, suggesting that surface features were relevant as a subsidiary process to the logic. Equally, however, because it is based on models containing entities, the theory inherits the ability to deal neatly with the widely noted presence and resilience of Illicit Conversion (see Chapter 5)

The other focus for skill acquisition obviously lies in the area of candidate conclusion testing, which can be subdivided into three kinds of tests: syntactic, pragmatic and semantic, where the last two can be distinguished only on the assumption that the subject has some kind of grasp of what it means for something to follow from something else. However, it has already been argued that this can be assumed of any subject who is able to make any sense of the task, since its absence reduces the exercise to a shambles.

The syntactic constraints require that the candidate conclusion must be a sentence. However, the ability to recognise sentences can be assumed to be present and stable (and usually effortless) in fluent speakers, so there is no scope for learning here.

The pragmatic constraints require that the conclusion must be in syllogistic form, and link the two end terms. These are explicitly described to the subject at the start of the experiment, though often they are not properly grasped – both Johnson-Laird and Inder have found subjects giving conclusions in "non-syllogistic" form or including the middle term. However, the results of Subject B (Section 7.4.2) suggest that both these types of errors die out[14] – the relevant filtering skills are acquired. This learning is to be expected because subject are able to distinguish appropriate from inappropriate conclusions, even though initially they do not grasp the importance of the difference. Subjects are being bombarded with examples of all the acceptable forms, so will be increasingly able to distinguish syllogistic from non-syllogistic forms. Grasping that the middle term should not be used results from the subject applying a Gricean criterion of relevance to the premises – since two premises are provided, both ought to be relevant to the conclusion. This can also be seen as an application of the exam skill of recognising that the questions rarely contain

[14] Subject A's only garbled response was early, too. However, there were one or two exceptions – the subjects rejected because they did not manage to recognise these constraints at all.

irrelevant information. Since the desired conclusions make "tidier" answers to the questions, and the distinguishing criteria are both available and simple, most subjects have no trouble in acquiring effective pragmatic filters.

The logical testing skills are the most difficult to acquire, principally because the subject has no feedback, and cannot recognise the correct conclusions. As has been argued above, subjects already have an idea of what it means to say that something must be true if something else is. They will also be familiar with the idea of trying to construct a situation in which a "consequence" does not follow, though from a situation completely unconnected with syllogistic reasoning. Everybody has made up excuses. Excuses are necessary when a person will (or did) not do something, even though the fact that they should (have) is shared knowledge between them and someone else. Inventing an excuse consists of trying to imagine another state of affairs that is consistent with everything that other person knows, but in which the obligation doesn't exist or can't be satisfied (through no fault of the defaulter). In syllogistic parlance, making an excuse could be seen[15] as trying to find a counter-model to the deduction:

Y is able to do X
Y knows Y should/must do X
Therefore
Y does X

Although the basic framework of the task is familiar, Johnson-Laird points out that there are any number of ways of modifying a model that are totally irrelevant to the truth of any given sentence. Recognising those ways that are, or even might be, relevant in a particular situation is a skill. However, it is in just this area that subjects' experience of similar tasks is crucially different from syllogism solving. The search for what kind of excuse to make would be controlled by the subjects background knowledge (of what is plausible), the influence of which syllogism experiments are designed to minimise. Moreover, most people have no concept of carrying out an exhaustive (or even methodical) search for an excuse. Finally, making excuses is basically a matter of bringing in new information to undermine one of the "premises" of the above "deduction" – showing how the obligation was not grasped or circumstances conspired to make the task impossible. But adding new information or denying the premises are not allowed in syllogistic reasoning! Thus although what is required is grasped and the overall strategy for a solution is familiar, the crucial detailed skills necessary to achieve it – to tell what model modifications are relevant at any particular time – will be missing.

Finally, because the theory envisages subjects attempting to find some way of tackling the unfamiliar problem of considering the truth of set relations, it can be easily modified slightly to accord well with the reported introspections of (virtually all of) the Edinburgh subjects[16]. Interviewed as soon as they had finished the experiment, some subjects did indeed say that they started out by imagining groups of people, but most mentioned imagining intersecting circles, or

[15] Although there is no suggestion that this formulation is in any way relevant to the way the process is normally carried out.

[16] It requires no modification at all to deal with the common feeling that syllogisms are an unfamiliar, formal problem-solving exercise!

derivatives of Venn diagrams. Indeed, one subject actually seemed to have invented her own notation, featuring circles linked by blocked or unblocked lines. With Johnson-Laird's account, all these reported intuitions must be dismissed as epiphenomenal. There is no place for diagrams or circles in his account, since they are not up to the tasks required of the (conceptual) mental models. However, this is something of a disaster for him, since his whole theory was inspired by the introspections of a subject about imagining groups of people!

According to the new theory, the subjects could well be constructing **and really using** something akin to Euler diagrams. They would, as a result of suspension of belief induced by the experimental situation, give up attempting to build and maintain models of "real" artists, beekeepers or whatever. Instead, they would try to use what they could remember of a formal notation that they had been shown in the past – Euler Circles. For, although none of the subjects had received any formal logic training, most children (nowadays) meet these techniques in mathematics 'O' level.

It must be stressed that there is no suggestion that capturing the situation by means of intersecting circles is a "natural" way of tackling reasoning tasks, nor in any way applicable to (informal) everyday thought. Any Euler circles "drawn" by the subjects would be consciously produced mental images (or perhaps models of pieces of paper with circles on). This is in sharp contrast to other theories that have implicated such diagrams in the solution of syllogisms (e.g. Erickson, 1974), which have assumed that they represent a significant feature of the solution mechanism. As a result, no simple correlation is predicted between numbers of diagrams employed and difficulty of solution. Although this number might correlate with the complexity of the problem in some way, this would not by any means be the limiting factor in subjects performance. The current theory asserts that if subjects use these diagrams, they do so consciously, because they have decided that something that they were explicitly taught in maths is of use.

What is more, the notion remains in line with the underlying assumption that subjects are using a mental model, although it is obviously not one of a group of individuals. Instead, they are creating an image of a number of intersecting circles (or, alternatively, a model of a plain surface with the circles drawn on it) which they then consciously interpret using the rules and techniques they were taught at school. It is easy to motivate such a suggestion, if only on the grounds that such a representation would have the advantage that, being (visually) much simpler, and easier to create, hold and manipulate. This is born out in the words of one of the subjects in Edinburgh, when asked whether they attempted to imagine people, I did at first, but I realised that they were only getting in the way, so I gave up.

However, such an imagined arrangement of circles is still a mental model, and as such is subject to a number of constraints. Specifically, subjects can only entertain one arrangement of circles at any time, and they must have definite positions, which means that any two circles either do or do not overlap. Thus although it may not be obvious, this superficially quite distinct suggestion is, in fact, practically identical to the previous account in terms of mental models of imaginary entities. Because the models discussed in the exposition of the account of syllogism solving only represent certain properties of well-defined situations, each is always precisely

isomorphic to a diagram involving circles. Equally, any procedure for manipulating models by creating or destroying entities can be directly expressed in terms of creating or removing overlap between pairs of circles. This means that each argument given above in terms of collections of entities is directly equivalent to one that accords directly with subjects' introspections, as indeed are the models of specific subjects constructed by using the computer program.[17] Thus, for instance, within an entity-based model Illicit Conversion is characterised by the omission of an individual while in the equivalent "imagined circles" account it arises by the selection of one rather than another of the two possible representations of the premise. This is in complete contrast to Johnson_Laird's models, which are enriched with the direct representation of vagueness and negation, and thus cannot be simply mapped into any kind of Euler circle.

7.5. General Features, Limitations and Further Work

There is a qualitative differences between the kinds of models of subjects that this theory is proposing and those that Johnson-Laird, and indeed most other theorists, have put forward. They attempt to capture the "average" subject, and produce an explanation for the statistical distribution of responses that are produced. This objective is based on the assumption that syllogistic reasoning is an instance of, or at least akin to, an important, natural mental activity – i.e. thinking! This leads them to assume there is a single common mechanism (or possibly a small number of them) available to all subjects and used by them for tackling syllogistic reasoning problems. As a result, subjects are expected to fall victim to the same kinds of errors, and since the mechanism must allow human beings the *competence* to reason soundly, these must be *execution* errors. From these assumptions, they hypothesise a mechanism which will malfunction in the right ways and produce predictions of the range in which the responses given by a subject (any subject) are likely to fall, possibly even with associated probabilities.

In contrast, the new approach is based on the assumption that syllogistic reasoning is an unfamiliar task, regarded by subjects as akin to an exercise in consciously grappling with a problem. From this viewpoint, subjects can be expected to approach the task in the light of the ways of solving problems that each has built up over a whole (unique) lifetime. There is no reson to expect them to begin by using the same methods or be prone to the same errors. More importantly the procedures they adopt have not been perfected by continuous application to everyday situations, there is no reason to expect them to be either sound or stable. Indeed, quite to the contrary, there is every reason to expect these new skills to develop as the (intelligent, typically undergraduate) subject practices this unfamiliar task. Thus there is no temptation to fall for what Newell (1981) termed the *Fixed Method Fallacy*, and there is every reason to doubt the worth of predictions based on data from groups of subjects – detailed knowledge of the growth patterns of the average child are not adequate for getting a school uniform that fits. Instead, the new approach

[17] The program was written early in the development of the theory, before subjects' introspections were available to indicate the prevalence of set diagrams. At that stage it was important to demonstrate how the meanings of the syllogistic quantifiers could be captured procedurally, and to illustrate the kinds of procedures that might be involved in linguistic access to non-lexicalised models. It would, in fact, be a trivial task to re-code the program to operate in terms of overlapping circles, but since this would not affect its performance, it has been left (and thus reported) as it was originally written.

attempts to capture the specific mental machinations of individual subjects, which are characterised in terms of the collections of heuristics that they are postulated to be using.

Because the theory accepts the notion that mental models are a widely used form of information structure, the representations it uses are constrained to satisfy the restrictions that this imposes. However, since no similar restrictions have been offered on the number or kind of strategies that subjects can employ, there would appear to be a sense in which the new account is somewhat lacking in predictive power. Given the choice of an arbitrary number of unconstrained rules it is possible to capture any pattern of responses whatsoever – all that is required is a collection of 64 rules, one for each syllogism form. In fact, since learning is allowed, it is even possible to achieve the same effect by using only a single rule which modified after each item, so the theory can actually account for any piece of behaviour in at least two different ways!

The obvious reaction to such a situation is to attempt to impose conditions on what can count as an acceptable conclusion generation or falsification strategy. However, before doing so it is important to realise that since the particular task at hand is a very specific and unfamiliar one, it is therefore highly unlikely to have (had) any bearing on the kinds of procedures that can be mustered to tackle it. Instead, the range of techniques that subjects are able to use will be constrained by the mental apparatus and abilities that they bring to the problem. This means that any restrictions upon them that are imposed must be influenced not by the desire to explain syllogistic reasoning nicely, but by the consideration of the common features of a wide range of phenomena. However, the study of such things is nothing less that the inter-domain theory that Newell (1981) complains we have not got.

Newell's own candidate for a task independent theory, based on the idea of the *problem space*, was described in Section 5.4.4. There, it was suggested that at least his application of it to syllogistic reasoning was completely inappropriate, since it assumed that subjects had a firm grasp of the the situation and were able to adopt a suitable representation of the problem space. Unfortunately, such a notion is completely at odds with the level of performance typically observed from naive experimental subjects. However, Newell stresses that this approach is only suggestive, and that he is mainly concerned with the application of the idea of space searching. Therefore, it is worth considering whether, once it is recognised that subjects find the task far from "simple and transparent", their performance may be captured in terms of searching some other problem space.

The kinds of strategies that have been employed in subject modelling are at least in line with the kind of ideas that Newell is proposing. Subjects are indeed seen as attempting to search a possibility space, although their poor grasp of the structure of the problem means that they often lack the necessary operators to complete the search. However, the subjects' level of ability, at least in terms of getting valid responses, means that the task is at neither of the extremes – triviality or accomplished skill – where Newell believes that space searching will be apparent. It is firmly in the "muddle in the middle", where the effects he hopes for might well be masked by subject's ill-conceived ideas of what they are doing. This is certainly the picture suggested by the modelling of individual subjects, where although their overall grasp of the problem was sound, their knowledge of how to go about solving it was lacking. Worse still is the fact that subjects are learning.

Newell's approach is based on the assumption that a fixed set of operators – either domain independent ones or those of an established skill – are being used to search a space. However, the results described above strongly suggest that subjects' performance – their range of operators – is continually changing. As a result, even though the idea of the problem space might be a valid and powerful inter-domain theory, and thus sorely needed, its power to restrict theories of syllogistic reasoning is likely to be minimal.

The likely inability of Newell's approach to impose any firm restrictions on subjects' behaviour should not come as a great surprise when one considers the range of approaches to syllogistic reasoning. At one extreme is expertise – the state reached by experimenters in syllogistic reasoning processing subjects' responses, who are simply able to recognise the appropriate conclusion to any pair of premises. Performance at this level is simply a matter of utilising an established direct mapping to the appropriate conclusion. At the other extreme, subjects are undeniably open to influence by the fact that their uncle George is a beekeeping chemist – they can take account of the meanings of the premises and their knowledge of the world. Moreover, there is no reason to doubt that some subjects are able to intermix both types of operations. As a result, if there are any limits on the kinds of procedures that subjects **can** deploy, they are broad – as broad as the range of resources that an adult cognitive system has at its disposal.

Fortunately, the fact that subjects can do almost anything to solve syllogism does not mean they do. In the absence of a way of delimiting what procedures subjects can utilise, it is a reasonable research strategy to attempt to discover those that they actually do employ. The first step on such a route is the modelling of a number of subjects, constraining the rules hypothesized purely intuitively, and attempting to interpret what this reveals. The models reported in Section 7.3 have been constructed in this manner. Rules were formulated roughly in line with the idea of a production system, in that they took the form of rules for actions and patterns to specify their applicability. Of course, this still leaves enormous freedom, and the constraints used to evaluate one possible model against another are worthy of attention.

The constraint with the most sweeping effect was that any strategy proposed should be matched against every permutation of the premises and the terms within each. This has the effect of preventing the use of very specific patterns for rules, which could otherwise be set up to generate any odd responses. However, it also seems to be necessary to the credibility of the proposed psychological processes of the subject, since without it subjects would be modelled as only ever considering the the premises precisely as they appear. Considering that they remain in full view in front of the subject at all times, it seems unlikely that they never take advantage of this. Unfortunately, the implementation of this feature within the simulator is not without cost, since it ensures that the permutation of the premises is always done in a fixed order, whereas it seems likely that, at least in the symmetrical figures, subject's behaviour is somewhat less regimented.

The other dominant criterion for judging a model is total number of active rules and learning events (as pointed out above, these can be traded off against each other). This is obviously

justified simply in terms of minimising the number of free parameters used to fit the data (rather than the number potentially available, for which no particular limit seems justifiable). The models that have been produced use approximately 6 such rules/learnings per model in order to achieve more than 60 correct responses (out of 64). In addition, learning events have been constrained as much as possible to correspond with long thinking times, although this is not a rigid connection – a flash of inspiration only takes a moment – and to occur only when appropriate material is being considered – i.e. it is undesirably to propose that subjects learn how to generate a conclusion from an **I** and an **O** premise while studying the **1AA** syllogism! Finally, the rules have been formulated in a manner that is as general as possible. Specifically, both generation and falsification strategies have been formulated in terms of re-arrangements of or associations between strings of words. This has the advantage of being unrelated to syllogistic reasoning, and thus not requiring any additional representational resources – subjects **must** have the ability to represent strings of words. Finally, and possibly least well defined – is the objective that the rules that subjects' are postulated to bring to bear (at least initially) must have a reasonably natural interpretation – they must be the kind of thing that a someone without logical training could be expected to think of!

Attempting to model specific subjects has shown that while it is comparatively easy to settle upon a reasonable model of a particular subject – one that is able to cover between 40 and 50 of the subject's responses, it is very much more difficult to refine that model further, to account for more than 60 conclusions, is **very** difficult. In addition, the space of successful models that satisfy these constraints is very sparse – that is, it has never been possible to see two (comparably acceptable on the above criteria) ways of producing any group of responses.

On the basis of the subjects modelled, it appears that most people have the *A-Effect* strategy for suggesting conclusions.[18] This single strategy is able to contribute towards both Johnson-Laird's Figural Effect and the traditional Atmosphere Effect. They usually also know how to defeat **A** conclusions, though seldom any others. Indeed, this is the ''core'' model that has been used as the starting point, as may have been suggested by the description of the modelling of Subject A, which also includes the range of responses that it gives. Interestingly, if this very simple package (one strategy of each kind and premise permutation) were proposed as an ad hoc stand-alone theory of syllogistic reasoning, it would predict the most common single response in Inder's experiment for 54 of the 64 syllogisms, and accounts for 54% of responses (these figures rise to 56 and 57% if it is allowed to predict both term orderings for reversible conclusions in symmetrical figures). These figures make it at least comparable to the majority of ''explanatory'' accounts of syllogistic reasoning that have been proposed. The other point to come out of these models is the high variability of subject's behaviour when dealing with combinations of an **E** premise with either an **I** or an **O**, with fluctuations in the handling of this class of problem being a common source of variability within subjects.

---

[18] This is the rule used to illustrate a typical conclusion generation strategies in Section 7.2: 'If one premise is of the form ''All A are B'', then try replacing any A in the other premise by B'.

The remaining variability between subjects is in the interpretation of the premises. In every case, subjects have been modelled as being fully aware of the truth conditions of the forms of sentences involved in syllogistic reasoning. However, subjects appear to differ, in the initial models that they build in response to the ambiguous premises – i.e. their tendency to illicitly convert the premises. Since this operation, like sentence validation, is a function of their language abilities, it has been assumed that subjects are consistent, both over time and between items, and that every model built is a situation in which both the premises hold.

While these "natural" criteria tightly constrain the range of models, the fact that the possible range of procedures is so wide makes it unlikely that it will be possible to **disprove** the theory by qualitatively failing to model a subject. As a result, it is tempting to consider other ways in which it might be directly falsified, one of which is suggested by its commitment to the notion that true execution and capacity errors are rare. Since the theory recognises the importance of the differences between individuals, it has never sought to be able to predict (or even restrict) the behaviour of "Mr. Average". However, when it was initially conceived, it was hoped that it would be able to achieve a completely different **type** of predictive power. The fact that no previous worker in the area had placed any emphasis on changes in subject's behaviour had given rise to an assumption. While it was recognised that each subject would be using a different set of reasoning strategies, it was expected that, at least for the duration of a single experiment during which they received no feedback, that set would be constant. This, together with the lack of performance errors, made it possible to hope that while the theory would be unable to say anything about an arbitrary subject it would, on the basis of a model built on the basis of some items, be able to predict precisely what responses would be given to others.

This notion has met with some support from the models for Subject C (described above), the other subject who gave very stable responses on his 4th and 5th sittings,[19] and a (logically trained but nonetheless error-prone) pilot subject. However, the bleak reality that became clear from attempting to model naive subjects on their first sitting was that their behaviour changed. Their responses simply could not be produced by a static package of strategies without removing the permutation-of-premises mechanism and introducing a plethora of specific procedures to compensate, the undesirability of which was mentioned above. Similarly, the attempt to model "Mr. Modal" – the notional subject who gave the most common response to every item – was also unsuccessful and, in the absence of any kind of "average ordering" (the presentation sequence having been randomised for each subject) had to be abandoned. Thus these attempts, which at least illustrate that intuitive restrictions on the strategies that can be employed have some power, serve primarily to highlight the importance of the learning effect.

It has been suggested, therefore, that rigid restrictions on acceptable strategies are unlikely to be forthcoming on a theoretical basis, and as a result the "natural" restrictions outlined above are

[19] The last two sittings of this subject were almost identical. They were also very nearly perfect demonstrations of logical competence. Two of the four differences between the two sittings were term order in reversible premises in symmetrical figures, while the other two were each an invalid conclusion given on only one of the two occasions (i.e. a single performance error in each batch).

likely to be the best that is available. Nevertheless, the goodness of fit of models built within these constraints is so striking that it seems well worth considering if there are any other possibilities for improving the falsifiability of the account.

The situation, and indeed the overall style of the theory, bear clear family resemblances to the work carried out by Brown, Burton, Friend and vanLehn on the analysis of children's subtraction skills in terms of interacting sub-skills. The core of their work is the explanation of the results of an individual subject, and they too employ a computer program to illustrate that the procedures they claim can give rise to the child's performance can indeed do so. Finally, they too begin from a situation of having only "intuitive" constraints on the subject models that they propose. As a result, it is worth examining the course of their work and considering its relevance to the proposed theory of syllogistic reasoning.

The initial stages of their work are reported in (Brown and Burton, 1978). They take knowledge of the differences between – i.e. the result of subtracting – single digits to be primitive, and see the procedure of subtracting two multi-digit numbers as the coordinated application of this knowledge. Subtraction is described as involving a number of distinct skills, such as knowing how to treat one column in a large subtraction or how to borrow. Learning how to subtract is the process of acquiring the relevant skills, but each may be imperfectly grasped, or "buggy", leaving the subject with a defective subtraction procedure. In analysing the subtraction performance of thousands of individual school children, Brown and Burton have observed approximately 100 "bugs". These are distinct defective versions of the skills required, such as always subtracting the smaller digit from the larger.[20] or always taking the result of a subtraction from zero to be zero. They then use a computer program, BUGGY, to demonstrate how the generally inaccurate performance of each individual child can be seen as resulting from the reliable execution of a particular set of strategies.

Because some pairs of bugs are incompatible, there are 3000 combinations possible, but only 300 of these are actually observed. Moreover, Brown and vanLehn point out that

> Using some Buggy ideas, Hetzel (private communication) detected that some students that were diagnosed as having a particular bug one day would appear to have an equally systematic but different bug a few days later (This often happened over a weekend where there was no intervening instruction). Since then we have discovered that many educators believe that the systematic errors they observe in a given student often spontaneously disappear and then reappear at later times.[21]
>
> (Brown and vanLehn, 1979. P6)

They try to explain these phenomena by suggesting that some bugs arise from attempts to repair others. They propose a model where defective procedures are deficient versions of the correct procedure, whether because of forgetting or incomplete or inaccurate learning. Such deficient procedures may lead to an impasse – a situation where no skill suggests a way to proceed. In these situations, they suggest that subjects invoke a repair mechanism that generates (possibly faulty) patches that will allow the solution to proceed, and that It is the set of all possible repairs to all

[20] The correct procedure, of course, is to always subtract the bottom digit from the top.

[21] Note that this is precisely the behaviour observed in Subject B.

possible impasses that is meant to characterise the set of all possible bugs (Brown and vanLehn, 1979. P9). Their results show that while the bugs that this approach predicts are indeed observed, they account for only 20% of those observed.

Burton (1981) extends the basic theory in another direction by attempting to automate the process of identifying the combination of bugs that describes each subject. He describes two computer program. The first, DEBUGGY, analyses a subject's performance on a subtraction test and produces the combination of recognised bugs that best explains it. In contrast, the second program, IDEBUGGY, attempts an interactive analysis of the subject's abilities, generating test items on the basis of earlier results in order to test for the presence of specific bugs.

There is clearly a great similarity between the initial stages of this project and the work just presented on modelling of syllogistic reasoning. However, there are a number of significant differences between the two domains that at least present impediments to their parallel development.

The first difference pertains to the volume of data, which in turn relates to the nature and importance of the task. Brown and Burton analysed the results of thousands of subtraction tests. This volume of data is only available because subtraction is being exercised and tested as part of the school curriculum. As a result, many hundreds of subject-hours can be freely committed to the test, since it is useful and close to what the children would be doing anyway. Further, a 20-item pencil and paper test is short and easy to administer to many people. In contrast, the current syllogistic test is longer[22] and best presented by direct interaction with a computer. Thus the cost (subject time devoted to the task) and effort (given the current state of computerisation of schools) of obtaining such a volume of data is much greater.

A second, and more fundamental, difference, is that the skill tested is one that the subjects are being explicitly **taught**. As a result, the end point of the learning process – the correct procedure that the subjects are working to acquire – is known. This makes it possible to break the task into separate skills that subjects will have to acquire, and this provides a skeleton onto which the buggy procedures can be hung. In contrast, syllogistic reasoning is being used to probe some kind of natural (thought or) problem solving processes. There is no a priori way of knowing what "correct" procedure they are working towards. Indeed, there is not even a way to be sure how their learning is directed at all, let alone whether it will ever produce a logically sound procedure. As a result, there is no way of identifying the skills that a subject must employ.

Finally, there is a gross difference in both the problem solving skills of the subjects and their relation to the task being examined. In contrast to school children at the stage of learning arithmetic, undergraduate students have at their disposal many years of not just experience, but active instruction in the ways and means of solving problems, and they represent the cream of a population, having been selected for their abilities to deploy just those skills. On the one hand are expert learners, to whom the specific (syllogistic) task is novel – they have never encountered it

[22] At least if all 64 syllogisms are presented. This is hard to avoid because of the difficulty of identifying related problems, which is involved with the second difference between the domains.

before the beginning of the test which is intended to analyse their skill. On the other hand, children learning arithmetic have undoubtedly struggled at length with the task prior to the particular performance that is being analysed. In the light of these facts, it is possible to make one undeniably obvious prediction: in stark contrast to children struggling with just one more batch of indescribable arithmetic problems, skilled undergraduate problem solvers can be expected to learn a lot from the task. Whereas Brown, Burton et. al. can expect to take a ''snap-shot'' of their subjects' subtraction abilities and hope to build up a library of components, the best that can be hoped for in first-time syllogism solving is a blurred image encompassing a progression of states.

This is a serious problem. It is, however, not insurmountable. Those who have attempted the task repeatedly have shown that subjects very quickly stabilise in their approach to the task. This suggests an experimental paradigm based around allowing subjects to solve syllogisms until they reached a stable performance profile, which the theory predicts should therefore be able to be modelled without learning. This would allow the specific strategies that each subject naturally adopts to be analysed clearly, without being obscured by the effects of learning. As a result, one could hope to build up a library of strategies which could then be analysed in order to extract the regularities in the rules that subjects are using.

Unfortunately, it is a task of daunting proportions, using a great deal of subject (and presumably experimenter) time to gather each piece of data. However, there is a particular feature of the experimental task that suggests that it may be possible greatly to reduce this cost, and at the same time test the theory in a completely different way. The feature of subjects' behaviour that has so far been regarded as the most troublesome – namely, the changes that occur the way that they tackle the task – can possibly be turned to advantage. Within the current theory, this learning is characterised in terms of the acquisition or deletion of strategies, which has been postulated to occur as a result of mechanisms that are beyond the scope of the current theory. However, one of the features of the syllogistic reasoning task that has cast doubt on the possibility of constraining the procedures that subjects employ also offers an ideal way of probing these otherwise inscrutable mechanisms. There is ample evidence that subjects are aware of lexical content of the premises, and can be influenced by it. It thus seems likely that the deliberate manipulation of this factor by the appropriate inclusion of strongly evocative material could serve to bring subjects' attention to particular defects in their reasoning. For instance, while the most common response to the **1AI** syllogism is an (invalid) **I** conclusion, arranging the premises so that the conclusion came out in terms of dogs and cats would seem very likely to disrupt this. Such disruption would be predicted to greatly increase the chances of the subjects improving their performance, which in the terms of the current theory, this would appear as immediate and appropriate (i.e. related to the anomalous item) strategy changes.

However, such manipulation of the material is not straightforward – the effect of the anomalous material would be the result of an interaction between the material itself and the way the subject was already addressing the problem. As a result, simply introducing semantically loaded material at various points among the test items may not yield as clear-cut results as would be desired. Ideally, it would be used to deliberately provoke specific strategy changes, although

this would require that it was introduced in the light of a model of the state of the subject's reasoning. This could be done either by the experimenter or, ideally, by a system along the lines of IDEBUGGY.

The task of modelling a subject is effectively a search among the space of possible models for one which gives a good fit to the observed behaviour. It is currently a skilled process, since there are no formalised rules for indicating the relative worth of any particular change to the model. However, the information that is employed is tightly restricted, and while the criteria used to evaluate which strategy changes are most likely to improve the situation are currently only a matter of intuition, there is no reason why at least approximations to them could not be formalised. As a result it seems reasonable to suggest that the task of modelling a subject is in many respects similar to many that are currently being tackled by expert systems, and DEBUGGY stands as an existence proof that something in this area is at least feasible.

It is therefore possible to imagine an experiment in which a subject's responses are analysed by machine (possibly with guidance from the experimenter) in order to dynamically create a model of the strategies being employed, and on the basis of this model, obviously anomalous test items would be introduced in order to deliberately change the subject's reasoning behaviour. Proposing such an experiment marks a significant change in the use of the task. It is no longer being proposed as a way of probing the processes of everyday thought, but as a tool for

7.6. Summary

This chapter has presented a new theory to account for the behaviour of subjects without logical training as they attempt to solve problems in syllogistic reasoning. It is unusual, in that it treats the task as an unfamiliar exercise in problem solving, and quite unique in the way if focuses on capturing the responses of an individual subject. Together, these features allow it to account for the common production of unsound conclusions in terms of *ability* (rather than *capacity* or *execution*) errors. Such an interpretation not only avoids having to suggest that most people cannot think about a problem for as long as a minute without making some kind of slip, but is also supported by the striking consistency of individual subjects. Moreover, the way that the surface features of the premises influence the production of (possible) conclusions allows it to offer highly plausible explanations of many of their observed effects on responses. Finally, the cognitive architecture and information structures it employs are in line with the those (and the restrictions on them) proposed in the early chapters of this thesis.

CHAPTER 8

Summary and Conclusions

8.1. Summary

The ultimate objective of cognitive psychology is to create and demonstrate an understanding of the functioning of the human mind. Doing so involves being able to describe it, to predict it and to explain it in terms of the interaction of familiar and established ideas. The approach that is normally referred to as the computational view of mind is based on the assumption that the most promising source of these ideas is computation. Indeed, many believe it offers the only conceptual framework that is remotely suitable for discussing and understanding the complexity of the phenomena of mind and the independence of their personal-level descriptions from the physical structures that support them.

One of the most explicit expositions of the implications of this approach, that of Pylyshyn, was presented in the Introduction. He argues that every computational process must be supported by some kind of "computer" – a fixed mechanism that is so structured that its deterministic operation is able to exhibit the flexibility that is associated with both computation and mental phenomena. As a result, he suggests that the principle objective of cognitive psychology should be the characterisation of the cognitive architecture that underlies human mentation. Unfortunately, Pylyshyn's theories are not without problems. In particular, the Introduction also argued that the notion of computation that he advances, in terms of the manipulation of representations in line with **explicitly represented rules**, does not pick out a well defined class of systems. As a result, an alternative concept was proposed as offering a characterisation of the features of mind on which the study of computational systems might shed some light. A *Flexible Autonomous System* was defined as a system that can modify some physical features that are causally involved in directing its behaviour. However, this has the effect of undermining a number of the consequences for psychology that Pylyshyn attempts to draw, particularly in connection with the possibility of recognising and seeing through an emulated architecture. Nevertheless, the idea that the cognitive system must be supported by an architecture that is beyond its ability to modify remains, as does the central position its identification occupies within the objectives of psychology.

Since the falsifiability of a theory is a measure of its value, sound methodology directs the psychologist to strive always to postulate the weakest possible architecture – that is, the one that will be unable to do the largest number of things. For instance, Pylyshyn suggests that certain interpretations of mental imagery propose that it reveals the operation of mechanisms for mental activity that have very severe limitations on the kinds of behaviour that they can produce. In

particular, he has in mind the common interpretation of imagery in terms of the manipulation of two-dimensional arrangements of information about the properties of object surfaces. However, Chapter 1 argued that such approaches have serious limitations both on their desirability and, as Pylyshyn himself showed, their implications for the cognitive architecture.

Nevertheless, the idea that imagery may be the manifestation of the manipulation of a specific type of perception-related representation still has the potential to restrict the cognitive architecture. To explain the operation of a system, and thus a fortiori the cognitive system, it is necessary to break its operation down into the interaction of a number of (simpler) sub-systems. In Chapter 1 Fodor's notion of an input module, which constitutes an argument for precisely such a subdivision of the sensory systems, was presented, and Chapter 2 proposed its extension to the mechanisms of behaviour production and even specialised thought processes. Such a modular architecture is the ideal choice for meeting the methodological objective of explaining behaviour using the weakest possible mechanism. The human mind is able to deploy processing resources that are capable of representing and manipulating literally any conceivable thought. Obviously, any satisfactory account of its behaviour must postulate mechanisms of sufficient flexibility and power to explain this. The importance of a modular architecture is that, by definition, it is able to restrict the range of phenomena to which that power can be applied. The black box nature of modules, and their informational encapsulation, can prevent resources that could facilitate the solution of a particular problem being brought to bear. As a result, a class of problem can prove intractable for a modular system even though it has the means to deal with it because they are located, and thus locked, within the inaccessible interior of a module.

The inabilities of a modular system arise from the restriction on the flow of information between modules, and the most effective way of weakening such an architecture (without denying it the power to approach the peaks of human abilities) is to further limit this communication. As a result, Chapter 2 also proposed, following the model of the *blackboard* architecture for expert systems, that the module interactions that give rise to cognitive phenomena are restricted to the state of a single (shared) information structure. This is obviously in keeping with the ideas of Johnson-Laird (1983, Chapter 16), who also describes cognitive behaviour as a manifestation of the interaction of a number of simultaneously active processing systems, although he says little about any limitations on their inter-communications.

The key feature of this shared information structure is that it is limited in what it can represent. In particular, it was suggested that it is restricted to a single possible state of the world, which led to it being dubbed a *mental model.* Moreover, the fact that it is the medium for all inter-module communication means that any restrictions upon it will affect all areas of mental activity. Such a universal influence is, of course, precisely what leads Pylyshyn to argue for the importance to psychology of the determination of the functional architecture that the brain provides to support mental processes. Fortunately, as he points out, it also allows results from all areas of psychology to be interpreted as bearing on its capacities.

It was suggested in Chapter 2 that imagery should be recognised as the manipulation of a mental model by a non-cognitive process – that is, one that is outside the Flexible Autonomous

System that underlies cognition. If this is so, then the limitations on the content of images – which are of necessity much looser than those that Pylyshyn attempted to impose – can be taken to be indicative of similar limitations on the model itself. Unfortunately, the richness of mental phenomena forces the recognition that the system that supports them is very powerful indeed. In particular, the fact that images can be scanned or "zoomed in on" reveals that the restrictions of this limited capacity are ameliorated by a powerful system which many features of the phenomena suggest is highly knowledge-intensive. Nevertheless, given the methodological desirability of doing so, it is possible to interpret certain psychological phenomena as evidence of significant restrictions on it. In particular, it was suggested that a mental model could not directly represent negation or ambiguity.

In the same vein, Chapter 3 attempts to apply these imagery-motivated restrictions to language by following Johnson-Laird and suggesting that language comprehension necessarily involves building a model of the situation under discussion. This implies that the familiar discourse model ought to share the restrictions on ambiguity and negation imposed on mental models. This means that since every linguistic statement is ambiguous – it specifies only a small subset of the properties of a situation – incorporating it within a model requires making (knowledgeable, but nonetheless uncertain) assumptions. As a result, the need to achieve understanding through modelling gives rise to bridging inferences, and shapes discourse (in terms of coherence).

However, while it is (comparatively) easy to see how making a mental model can contribute to the understanding of a specific situation, human mental life also includes, and even features, the ability to deal with abstract situations and ideas. The particular abstract task that Johnson-Laird selected for consideration, which was described in Chapter 4, is solving syllogistic reasoning problems. This is an activity which, as Chapter 5 describes, had already been the subject of a number of experimental investigations, and given rise to a number of psychological theories that make no mention of any kind of mental model. In order to show that the mental model is of central importance to the understanding of mental activity, it is obviously essential to suggest what role it plays in solving such problems within a theory that is at least as good as any that have been proposed before. As usual, sound methodology dictates that it should be the weakest mechanism capable of performing the task. Ideally it should also complement the use of a mental model for concrete situations, and certainly it should not undermine it, in the sense of providing a mechanism that is better suited to dealing with specific tasks than the models themselves.

The most obvious way of deploying mechanisms geared towards specific situations in order to deal with abstractions is to use them to represent exemplars or typical instances. This is a notion which can be traced back at least to Hume and Berkeley, who pointed out how much forming a suitable image can facilitate geometrical reasoning – an observation which even Fodor (1975, P192-3) is forced to admit does suggest that "(some) internal representations are, or may be, nondiscursive". Hume recognised that it is not possible to form an image of a triangle in general, but only of a particular one, which opens up the possibility of drawing conclusions on the basis of properties peculiar to the chosen example. However, he believed that if any such conclusion were drawn, then suitable counter-examples would readily come to mind to highlight its

inappropriateness.

Johnson-Laird (1983, P157) agrees with Hume's underlying suggestion of reasoning with an exemplar and then checking for over-generalisations. However, he differs to the extent that he stresses that the reasoner must "take pains" to avoid drawing a conclusion that hold only for the specific instance considered – i.e. must actively search to ensure that it does not rely on properties unique to the chosen exemplar. Indeed, this is the basis of Johnson-Laird's proposal for the solution of syllogistic reasoning problems, which is presented in Chapter 6. This involves testing candidate conclusions by considering a number of the situations that are compatible with the premises.

One of the most striking features of syllogistic reasoning problems is their range of difficulties – they vary from trivially easy to virtually impossible. Indeed, as pointed out in Chapter 4, accounting for the patterns of subjects' reasoning errors is one of the most important tasks for any theory of performance on the task. Most recent work (outlined in chapter 5) assumes that people have the competence and ability to reason correctly. That is, since some logicians can reason well, and everybody can recognise faulty reasoning (at least when it is pointed out), people must have some kind of sound reasoning procedure available. This means that the numerous reasoning errors that are observed in experiments must reveal some problem with the execution of that sound procedure. In the terminology introduced in Chapter 4, they are interpreted as *execution* or *capacity* errors, and thus assumed to be the manifestation of some kind of limit on the cognitive architecture

Johnson-Laird's account follows in this tradition, since he too provides a mechanism for solving syllogisms correctly and characterises reasoning errors in terms of its malfunctions due to performance factors. This obliges him to ensure that (common) errors can arise from "simple" faults – i.e. specific omissions or corruptions. This amounts to a requirement for a representation for situations that has explicit negative information and qualitative uncertainty – he has to use *conceptual*, as opposed to *physical*, mental models. It was suggested, in chapter 6, that this is the most significant misfeature of his account, since it constitutes a considerable extension of the power of the cognitive architecture.

In contrast, the account of syllogism solving presented in Chapter 7 uses only resources that are known to be available. It postulates only the ability to represent deterministic situations (in line with the maximal constraints proposed previously) and to represent and understand simple English sentences. These resources are adequate because the account rests on the fundamental assumption that subjects' incorrect reasoning is caused by *ability* errors. This amounts to a denial that subjects are only a slip away from perfect performance, and an assertion that, in the absence of the skills necessary for correct reasoning, they are simply muddling by as best they know how. From such a viewpoint, there is no reason to expect that people will tackle the problems in the same way, or that anything meaningful will be revealed by an analysis of the average performance of groups of subjects. Instead, it seems appropriate to focus on the modelling of individual subjects, and the strategies that they bring to bear, in the hope that this will reveal something about the kind of problem solving resources that those individuals have available. Such an approach has the

disadvantage that it is not possible to use statistical methods to reveal trends among noisy results, but the results of the subject modelling reported in chapter 7 suggests that this is not a problem.

Finally, Chapter 7 also outlined the results from applying the new theory of syllogistic reasoning. These showed that the reasoning process can be captured in a way which suggests that it is the result of the reliable operation of a deficient mechanism – that is, there are very few performance errors. In addition, they served to highlight the extent to which subjects' behaviour changes during the course of even a single syllogistic reasoning experiment. As a result of this it was suggested that while syllogistic reasoning is not a manifestation of the processes that support everyday thought, the experimental paradigm has the potential to allow exploration of the deeper processes associated with learning.

8.2. Conclusions

Evidence from neurology reveals the slowness of the brain's fundamental processing units, which are mostly only capable of considerably less than 1000 operations per second, or fewer than 200 during the time it takes to shadow a sentence. This clearly emphasises the need to recognise the importance of its ability to process large amounts of information in parallel. Similarly, it imposes considerable restriction on the distribution of information about the brain – neurons are simply not fast enough to implement any kind of "central store". It seems that the observed performance can only be seen as compatible with the mechanisms that support it if it is conceptualised as arising from a number of concurrently operating, fairly self-contained, processing systems. This certainly offers a natural explanation of why the solution to a difficult problem so often occurs in dreams or in the bath – or indeed in any of a host of situations in which the problem is far from the focus of attention. Similarly, when Johnson-Laird asks "Why are there silences when we think aloud", it prompts the reply that at those times, there is simply no **cognitive** processing going on – the subject's cognitive processes are simply awaiting the deliverances of some other module. Indeed, the activities of those modules seem to naturally provide an engine for the psychoanalysts' notion of the subconscious, and a place for the advertising man's seeds to lodge.

Recognising that cognitive mental life emerges from the operation of a number of independent processing systems offers a natural explanation for many of its "gross" anomalies. Fodor (1983) focussed on the impenetrability of perceptual skills, which serves to highlight that the property is shared by their behavioural counterparts, and the fact that people have no active access to the details of so many of the activities they carry out all the time – indeed, this inability constitutes the central obstacle to constructing an expert system.

However, equally striking are the cumbersome and indirect techniques needed for transferring a skill – it involves so much more than just saying what is to be done or corrected. Furthermore, the detrimental effects of fear or "trying too hard" result from the difficulty of suppressing the unwanted operation of a module. Finally, while the body – a physical object that has obvious restrictions on its ability to satisfy instructions – clearly constitutes a potential focus of contention between independent processing systems, there are other kinds of conflict. Specifically, the notion of the mental model as the single information structure that mediates all inter-module

communication gives rise to the possibility of an entirely informational resource conflict should multiple processes attempt to influence its state. In particular, selective task impairment is only to be expected – imagery will be facilitated by closing the eyes, since this will stop the visual input modules attempting to report their findings. Equally, this same feature of the architecture is able to account for the beneficial effect on the performance of a skill of imagining its execution.

There is, therefore, overwhelming evidence for the notion that conscious thought is not the only information processing carried out within the brain, which has, indeed, led to widespread acceptance of the notion. Nevertheless, the consequences of this, particularly arising from the need for communication and cooperation between these various systems, are seldom mentioned. When the normal function of the brain is disrupted, whether by surgery or injury, the ability of the cognitive system to continue functioning is astonishing. The mechanisms that allow the various processing systems to operate at all in the presence of drastic communication difficulties – for instance, those by dividing the halves of the brain – seem so powerful that it is hard to believe that they simply arose to cope with such a highly anomalous situation.

Perhaps the most striking observation is that, even though it is easy to demonstrate that severing the corpus callosum seriously disrupts inter-hemisphere communication, mental processing continues more or less normally. For instance, it is easy to show that the transfer of visual information from one side of the field of view to the part of the brain that controls the opposite side of the body is entirely disrupted. This means that the vocal (left) hemisphere is deprived of all information about the visual field to the left of the point of fixation. Yet Gazzaniga (1970, P72-73) describes "the peculiar phenomenon that split-brain patients do not complain about their inability to verbally describe visual information to the left of fixation." He points out that

> With intact half-brains like ours, it seems incomprehensible that we would not notice a difference in our visual world. Yet, split, the left hemisphere never complains, never alludes to a difficulty. It is as if the mechanism for the realisation that vision was once available across the midline exists only when the callosum is intact. With it gone, the left hemisphere simply doesn't respond to an appropriate question about this subject from the examiner. Indeed, one would miss the departure of a good friend more, apparently, than the left hemisphere misses the right
>
> (ibid)

Similarly Gazzaniga and Ledoux (1978, P157) report experiments that reveal that, in the monkey at least, the divided hemispheres spontaneously communicate through indirectly observing the activity of the other. These observations suggest that the system already has mechanisms for dealing with limitations on inter-module communications.

Some of the most astonishing features of the mechanisms adopted to deal with communication failures are revealed in the way that the left (verbal) hemisphere reacts to actions that the right hemisphere initiates on the basis of information to which it alone has access. For instance, when a stimulus is momentarily presented in the left visual field, the (left hemisphere of the) subject cannot give any verbal indication of what, or indeed whether anything, happened. Nevertheless, the right hemisphere is perfectly capable of using the left hand to give an appropriate response, such as by selecting a picture or object. In such situations, therefore, the left hand carries out an action for which the speech generation processes have no access to any explanation. However, this causes no distress – subjects express no concern that their left hand embarks on some inexplicable program of

activity. Instead, they simply invent an explanation for its activities that is coherent within the situation as the left hemisphere has perceived it. Indeed, the notion of *cognitive dissonance* as a mechanism for shaping an individual's beliefs, particularly about himself, can be interpreted as suggesting that such behaviour is a normal feature of mental life. So too can Evans' (1982) proposal that the processes that subjects use to explain their behaviour in reasoning experiments are unrelated to those that were actually employed.

Once it is recognised that information processing within the brain is organised as a number of independent processes, the importance of their intercommunication, and the influence that its limitations can have on behaviour, becomes apparent. Obviously, determining precisely what is communicated between modules is one of the principal goals of any psychology that accepts their existence. While there is (at least as yet) insufficient information to determine these details, it is still possible, and desirable, for more general theorising to proceed, provided a suitable "high level" description of this communication can be put forward. It has been suggested that a *mental model* fills precisely this role. It was also argued that since a model is a representation of a situation in a specific "code", there will be limitations on what a model will be able to represent and along what dimensions it will allow uncertainty. In particular, it was suggested that the model will be restricted to handling the properties of a single, highly specific situation. Moreover, since the survival value of the model would originally have been intimately connected with the coordination of sense information, this role may well have shaped the representations it employs. This in turn suggest that the sensory systems, rather than, for instance, language, will be the best source of insight into the structure of the information within a model.

Mention of linguistic behaviour serves to highlight that it is one of the most powerful devices available to the human cognitive system. It allows the use of a capacity that must be present in any theory of mental life – namely, the ability to represent a word – to cache almost anything in a model. All that is required is to add even a partial representation of the appropriate linguistic items, encoded as they would typically be if perceived. Of course, being able to do this does not constitute "having" the concept in any sense. The words appended to the model are simply uninterpreted syllable strings – there are no processing modules to manipulate or respond to them. They can only occupy a part of the "blackboard" and await some further linguistic processing. This may well take the form of (repeatedly) rewriting them, possibly in terms of a rote-memorised definition of the words involved, until this results in something that is properly understood – that is, something that there are processing modules to deal with.

This represents a familiar but previously uncharacterised state of understanding. The meanings of the terms and the facts of the situations involved are all known, but generating the consequences remains a slow, demanding and error-prone process. Indeed, it literally takes the form of conscious reasoning, in the sense of replacing one sentence with another that it implies. As such, of course, it will have all the properties and power of a formal logic, although unless the reasoner has been trained, there is no reason to expect that it will be either sound or complete.

There is even evidence to suggest that this kind of linguistic "marker" can be used for inter-module communication, even though language processing is widely known to be closely associated

with specific areas of the brain, normally in the left hemisphere. Studies of split-brain patients reveal that, although the ability to speak and to handle grammatical structures is localised to the left hemisphere, the right hemisphere is usually able to deal with a modest range of language use – see (Gazzaniga, 1970. P119 - 124) . Specifically, the non-verbal hemisphere is normally able to understand (select objects on the basis of) and use (name objects by pointing to) nouns and adjectives, and even spell simple words. However, there is normally no ability to deal with verbs, which leads and (Gazzaniga and Ledoux, 1978. P97) to offer the observation that

> Patients with dominant-hemisphere damage often suffer from either anomia or dysnomia. This inability to find a noun is never contrasted by observation of patients suffering from *averbia!* What is it about the role of the verb in the language process that makes this part of speech elusive in discrete brain disease? It seems that when verbs do go, the entire language system collapses... It is of interest to note that in the metalanguage training of aphasics, verbs were difficult to train, whereas symbols for noun-objects were generally learned in one trial.
>
> (Gazzaniga and Ledoux, 1978. P97)

However, these results certainly show that the ability to handle linguistic tokens, and particularly nouns, is distributed, which supports the notion that they could form the basis of some kind of inter-module communication.

The account of syllogistic reasoning presented in Chapter 7 illustrates this kind of linguistic processing. Understanding the individual sentences that constitute a syllogism is well within the abilities of the subject's language processing system, and can thus be accomplished correctly, effortlessly and reliably. However, actually producing a conclusion that follows from a pair of sentences is by no means a familiar task, and as a result no suitable procedures are available. Instead, the premises must be manipulated as uninterpreted (by the modules manipulating them) word strings until some kind of potential conclusion can be generated by rearranging their constituents. Once a suitable sentence has been produced, the language interpretation modules can quickly and effortlessly validate it against the current model of the situation described by the premises. The difficulty in the task comes in the shuffling of lexical items.

Although this kind of conscious, linguistic solution process may be slow and unreliable, it does provide an answer – it allows the system to do something in the unfamiliar situation. More importantly, it furnishes the ''blackboard'' with a sequence of examples of the behaviour required, which can be used to guide the development of a suitable processing module. Thus although the difficulty here arises in the combination of the sentences, rather than the concepts embodied in any one, the situation can still be used to illustrate the technique of manipulating strings of (English) symbols in order to support a crude solution process that will then serve as a focus for module development.

There is, then, a sense in which, despite Fodor's belief that it is impossible, English is being used as a ''Language of Thought'', although the way it is used is completely different from the kind of system that he had in mind. Crucially the functional architecture provided by the brain does not q understand the sentences – it is not structured to treat them appropriately – but instead they are manipulated without comprehension at all! Moreover, the language is not innate, but **learnt** – indeed in the way that steps are being taken in line with explicitly learnt inference rules, the system is practically an ideal example of Pylyshyn's notion of computation as rule following.

It is now scarcely plausible to deny that mental life is the result of many simultaneous information processing activities. Moreover, if it is to be understood, it must be in terms of the operation of a number of independently functioning, modular, subsystems. This means that the communications between these systems that allow them to interact and cooperate constitutes one of the tightest constraints on the behaviour of the overall system. It has been suggested that the blackboard architecture regulating the interaction of a number of independent, domain-specific processing modules offers the right kind of framework within which to attempt the task of explaining the phenomena of mental life. In particular, it has been suggested that the phenomena of mental imagery, and those that inspire the notion of a mental model, should be interpreted as revealing something of the properties of this central medium of communication.

References

Alty, J. L. and Coombs, M. J. (1984) *Expert Systems: Concepts and Examples.* Manchester: NCC.

Anderson, R. C. and Ortony, A. (1975) On putting apples into bottles - a problem of polysemy. *Cognitive Psychology*, **7**, 167-180.

Anderson, J. R. (1978) Arguments concerning representations for mental imagery. *Psychological Review*, **85**, 249-277.

Austin, J. L. (1962) *How To Do Things With Words.* Oxford: Clarendon Press.

Begg, I. and Denny, J. (1969) Empirical reconciliation of atmosphere and conversion interpretations of syllogistic reasoning errors. *Journal of Experimental Psychology*, **81**, 351-354.

Bower, T. G. R., Broughton, J. M. and Moore, M. K. (1971) The development of the object concept as manifested by changes in the tracking behaviour of infants between 7 and 20 weeks of age. *Journal of Experimental Child Psychology*, **11**, 182-193.

Bower, T. G. R. (1974) The evolution of the sensory systems. In Macleod, R. B. and Pisk, H. L. (eds.) *Perception: Essays in Honour of J. J. Gibson*, pp141-152. Ithaca: Cornell University Press.

Bower, T. G. R. (1982) *Development in Infancy*, 2nd Edition. San Fransisco: W. H. Freeman.

Brooks, L. (1968) Spatial and verbal components of the act of recall. *Canadian Journal of Psychology*, **22**, 349-368.

Brown, J. S. and Burton, R. R. (1978) Diagnostic models for procedural bugs in basic mathematics. *Cognitive Science*, **2**, 155-192.

Brown, J. S. and VanLehn, K. (1979) Towards a generative theory of "bugs". Cognitive and Instructional Science Series, Xerox Palo Alto Research Center, Palo Alto, Ca., December, 1979.

Bryant, P. E. and Trabasso, T. (1971) Transitive inference and memory in young children. *Nature*, **232**, 456-458.

Burke, J. (1985) *The Day the Universe Changed.* London: British Broadcasting Corporation.

Burton, R. R. (1981) Diagnosing Bugs in a Simple Procedural Skill. Cognitive and Instructional Science Series CIS-8, Xerox Palo Alto Research Center, Palo Alto, Ca., March, 1981.

Card, S. K., Moran, T. P. and Newell, A. (1981) *The Psychology of Human-Computer Interaction.* Hillsdale, N.J.: Lawrence Erlbaum Associates.

Carnap, R. (1947) *Meaning and Necessity.* Chicago, Illinois: The University of Chicago Press.

Ceraso, J. and Provitera, A. (1971) Sources of error in syllogistic reasoning. *Cognitive Psychology*, **2**, 400-410.

Chapman, I. J. and Chapman, J. P. (1959) Atmosphere effect re-examined. *Journal of Experimental Psychology*, **58**, 220-226.

Chomsky, N. (1965) *Aspects of the Theory of Syntax.* Cambridge, Mass.: MIT Press.

Chomsky, N. (1980) *Rules and Representations.* Oxford: Basil Blackwell.

Chomsky, N. (1980) On binding. *Linguistic Inquiry*, **11**, 1-46.

Chung, S. and McCloskey, J. (1983) On the interpretation of certain island facts in GPSG. *Linguistic Inquiry*, **14**, 704-713.

Clark, H. H. and Haviland, S. E. (1977) Comprehension and the given-new contract. In Freedle, R. O. (ed.) *Discourse Production and Comprehension*, Volume 1, pp1-40. Norwood, N.J.: Ablex.

Clocksin, W. F. and Mellish, C. S. (1981) *Programming in Prolog.* Berlin: Springer-Verlag.

Cohen, M. R. and Nagel, E. (1934) *An Introduction to Logic and Scientific Method.* New York: Harcourt Brace.

Collins, A. M. and Loftus, E. F. (1975) A Spreading Activation Theory of Semantic Processing. *Psychological Review*, **82**, 407-428.

Cooper, L. A. and Shepard, R. N. (1973) Chronometric studies of the rotation of mental images. In Chase, W. G. (ed.) *Visual Information Processing*. New York: Academic Press.

Cooper, L. A. and Shepard, R. N. (1984) Turning something over in the mind. *Scientific American*, **251**, 114-121.

Copi, I. (1968) Logic. In *Colliers Encyclopaedia*, Volume 14. Washington: Crowell-Collier.

Corteen, R. and Wood, B. (1972) Autonomic responses to shock-associated words in an unattended channel. *Journal of Experimental Psychology*, **94**, 308-313.

Day, R. H. (1969) *Human Perception*. Sydney: John Wiley and Sons.

DeGroot, A. D. (1965) *Thought and Choice in Chess*. The Hague: Mouton.

DeGroot, D. (1966) Perception and Memory versus Thought. In Kleinmuntz, B. (ed.) *Problem Solving: Research, Method and Theory*. London: Wiley.

Denis, M. and Carfantan, M. (1985) People's knowledge about images. *Cognition*, **20**, 49-60.

Dennett, D. C. (1969) *Content and Consciousness*. New York: Humanities Press.

Dennett, D. C. (1978) *Brainstorms: Philosophical Essays on Mind and Psychology*. Montgomery, Vermont: Bradford.

Dickstein, L. S. (1975) Effects of instructions and premise order on errors in syllogistic reasoning. *Journal of Experimental Psychology (Human Learning and Memory)*, **1**, 376-384.

Dickstein, L. S. (1978) Error processing in syllogistic reasoning. *Memory and Cognition*, **6**, 537-543.

Dickstein, L. S. (1978) The effect of figure on syllogistic reasoning. *Cognition*, **6**, 76-83.

Dreyfus, H. L. (1979) *What Computers Can't Do*. New York: Harper and Row.

Ehrlich, K. and Johnson-Laird, P. N. (1982) Spatial descriptions and referential continuity. *Journal of Verbal Learning and Verbal Behaviour*, **21**, 296-306.

Erickson, J. R. (1974) A set analysis theory of behaviour in formal syllogistic reasoning tasks. In Solso, R. (ed.) *Loyola Symposium on Cognition*, Volume 2. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Evans, J. S. B. T. (1982) *The Psychology of Deductive Reasoning*. London: Routledge and Kegan Paul.

Fahlman, S. E. (1979) *NETL: A System for Representing and Using Real-World Knowledge*. Cambridge, Mass.: MIT Press.

Fine, K. (1983) A Defence of Arbitrary Objects. In *Proceedings of the Aristotelian Society supp. vol. LVII*, 1983, pp55-77.

Finke, R. A. and Kosslyn, S. M. (1980) Mental Imagery Acuity in the Peripheral Visual Field. *Journal of Experimental Psychology (Human Perception and Performance)*, No. 6, 244-264.

Finke, R. A. and Pinker, S. (1982) Spontaneous mental image scanning in mental extrapolation.. *Journal of Experimental Psychology (Learning, Memory and Cognition)*, No. 8, 142-147.

Finke, R. A. and Pinker, S. (1983) Directional scanning in remembered visual patterns. *Journal of Experimental Psychology (Learning, Memory and Cognition)*, No. 9, 398-410.

Finke, R. A. (1986) Mental imagery and the visual system. *Scientific American*, **254**, 76-83.

Fodor, J. A. (1975) *The Language of Thought*. New York: Thomas Crewell.

Fodor, J. A. (1978) Tom Swift and his Procedural Grandmother. *Cognition*, No. 6, 229-247.

Fodor, J. A. (1980) Methodological solipsism considered as a research strategy in cognitive psychology. *Behavioural and Brain Sciences*, 3, 63-109.

Fodor, J. A. and Pylyshyn, Z. (1981) How direct is visual perception?. *Cognition*, **9**, 139-196.

Fodor, J. A. (1983) *Modularity of Mind.* MIT.

Forgy, C. L. and McDermott, J. (1979) OPS: a domain-independent production system language. In *Proceedings of the Sixth International Joint Conference on Artificial Intelligence*, Tokyo, Japan, August 20-23, 1979.

Frase, L. T. (1968) Associative factors in syllogistic reasoning. *Journal of Experimental Psychology*, **76**, 407-412.

Garnham, A. (1979) Instantiation of verbs. *Quarterly Journal of Experimental Psychology*, **31**, 207-214.

Gazzaniga, M. (1970) *The Bisected Brain.* New York: Appleton-Century-Crofts.

Gazzaniga, M. S. and LeDoux, J. (1978) *The Integrated Mind.* New York: Plenum Press.

Gibson, J. J. (1979) *The Ecological Approach to Visual Perception.* Boston, Mass.: Houghton Mifflin.

Grice, H. P. (1975) Logic and Conversation. In Cole, P. and Morgan, J. L. (eds.) *Syntax and Semantics*, Volume 3: *Speech Acts*, pp41-58. New York: Academic Press.

Hagert, G. (1984) Modelling mental models: Experiments in cognitive modelling of spatial reasoning. In *European Joint Conference on Artificial Intelligence*, 1984.

Hagert, G. (1985) What is a mental model? On conceptual models in reasoning with spatial descriptions. In *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, University of California at Los Angeles, Los Angeles, Ca., August 18-23, 1985, pp274-277.

Haviland, S. E. and Clark, H. H. (1974) What's new? Acquiring new information as a process in comprehension. *Journal of Verbal Learning and Verbal Behaviour*, **13**, 512-521.

Hayes-Roth, F. (1979) Distinguishing theories of representation: a critique of Anderson's "arguments concerning mental imagery". *Psychological Review*, **86**, 376-382.

Hebb, D. O. (1968) Concerning Imagery. *Psychological Review*, No. 75, 466-477.

Henik and Tzelgon (1982) Is 3 > 5: the relation between physical and semantic size in comparison tasks. *Memory and Cognition*, **10**, 389-395.

Henle, M. (1962) On the relation between logic and thinking. *Psychological Review*, **69**, 366-378.

Hinton, G. E. and Anderson, J. A. (eds.) (1981) *Parallel Models of Associative Memory*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Hollingdale, S. H. and Tootill, G. C. (1965) *Electronic Computers*. London: Pelican.

Hopfield, J. J. (1982) Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Science*, **79**, 2534-2558.

Hubel, D. H. and Weisel, T. N. (1962) Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology*, **160**, 106-154.

Hubel, D. H. and Weisel, T. N. (1968) Receptive fields and functional architecture of monkey striate cortex. *Journal of Physiology*, **195**, 215-243.

Hunter, W. S. (1913) The delayed reaction in animals and children. *Behavioural monographs*, **2**.

Hunter, I. M. L. (1957) The solving of three-term series problems. *British Journal of Psychology*, **48**, 286-298.

Inder, R. (1983) George II: A simulator that is not all it is claimed to be.. Working Paper No. 143, Department of Artificial Intelligence, University of Edinburgh, Edinburgh, Scotland, 1983.

Inder, R. (1986) Zen and the art of ecological perception: on and around J. J. Gibson's "Ecological Approach to Visual Perception". Discussion Paper, School of Epistemics, University of Edinburgh.

Inhelder, B. and Piaget, J. (1958) *The Growth of Logical Thinking from Childhood to Adolescence*. London: Routledge and Kegan Paul.

Johnson-Laird, P. N. and Wason, P. C. (1972) *Psychology of Reasoning: Structure and Content*. Cambridge, Mass.: Harvard University Press.

Johnson-Laird, P. H. (1975) Models of deduction. In Falmagne, R. J. (ed.) *Reasoning: Representation and Process in Children and Adults.* Hillsdale, N.J.: Lawrence Erlbaum Associates.

Johnson-Laird, P. N. and Steedman, M. J. (1978) The psychology of syllogisms. *Cognitive Psychology*, **10**, 64-99.

Johnson-Laird, P. N. (1983) *Mental Models.* Cambridge: Cambridge University Press.

Johnson-Laird, P. N. and Bara, B. G. (1984) Syllogistic Inference. *Cognition*, **16**, 1-61.

Kamp, H. (1981) A theory of truth and semantic representation. In Groenendijk, J. A. G., Janssen, T. M. V. and Stokhof, M. B. J. (eds.) *Formal Methods in the Study of Language*, Volume 136, pp277-322. Amsterdam: Mathematical Centre Tracts.

Kintsch, W. and Dijk, T. A. (1978) Towards a model of text comprehension and reproduction. *Psychological Review*, **85**, 363-394.

Knuth, D. E. (1973) *The Art of Computer Programming*, Volume 3: *Sorting and Searching.* Reading, Mass.: Addison-Wesley.

Kosslyn, S. M. (1975) Information representation in visual images. *Cognitive Psychology*, **7**, 341-370.

Kosslyn, S. M., Ball, T. M. and Reiser, B. J. (1978) Visual images preserve metric spatial information: evidence from studies of imagery scanning. *Journal of Experimental Psychology (Human Perception and Performance)*, No. 4, 47-60.

Kosslyn, S. M., Pinker, S., Smith, G. E. and Schwartz, S. P. (1979) On the demystification of mental imagery. *Behavioural and Brain Sciences*, **2**, 535-581.

Kosslyn, S. M. (1980) *Image and Mind.* Cambridge, MA: Harvard University Press.

Kosslyn, S. M. (1983) *Ghosts in the Mind's Machine.* New York: Norton.

Kuhn, T. (1970) *The Structure of Scientific Revolutions, 2nd Edition.* Chicago, Illinois: The University of Chicago Press.

Lackner, J. and Garrett, M. (1973) Resolving ambiguity: effects of biasing context in the unattended ear. *Cognition*, **1**, 359-372.

Lakoff, G. and Johnson, M. (1980) *The Metaphors We Live By.* Chicago, Illinois: The University of Chicago Press.

Lee, J. R. (1983a) The Nature of Johnson-Laird's ''Mental Models''. Discussion Paper, School of Epistemics, University of Edinburgh.

Lee, J. R. (1983b) Johnson-Laird's Mental Models: Two Problems. Discussion Paper, School of Epistemics, University of Edinburgh.

Lee, J. R. (1984) Johnson-Laird's models and truth. Discussion Paper, School of Epistemics, University of Edinburgh.

Levy, W. B., Anderson, J. A. and Lehmkuhle, S. (eds.) (1984) In Levy, W. B., Anderson, J. A. and Lehmkuhle, S. (eds.) *Synaptic Modification, Neuron Selectivity, and Nervous System Organization.* Hillsdale, N.J.: Lawrence Erlbaum Associates.

Lewis, J. (1970) Semantic processing of unattended messages using dichotic listening. *Journal of Experimental Psychology*, **85**, 225-228.

Luger, G. F., Wishart, J. G. and Bower, T. G. R. (1983) A model of development of the early infant object concept. *Perception*, **12**, 21-34.

Luria, A. R. (1977) *The Social History of Cognition.* Cambridge, Mass.: Harvard University Press.

Mani, K. and Johnson-Laird, P. N. (1982) The mental representation of spatial descriptions. *Memory and Cognition*, **10**, 81-87.

Marcus, M. P. (1977) A Theory of Syntactic Recognition for Natural Language. PhD Thesis, MIT.

Marr, D. (1982) *Vision.* San Francisco: Freeman.

Marslen-Wilson, W. D. (1973) Linguistic Structure and Speech Shadowing at Very Short Latencies. *Nature*, **244**, 522-523.

Marslen-Wilson, W. D. (1975) Sentence perception as an interactive parallel process. *Science*, **189**, 226-228.

McCloskey, M. (1983) Naive theories of motion. Chapter 13 in Gentner, D. and Stevens, A. L. (eds.) *Mental Models*, pp299-324. Hillsdale, N.J.: Lawrence Erlbaum Associates.

McGonigle, B. and Chalmers, M. (1977) Are monkeys logical?. *Nature*, **267**, 694-696.

Meltzoff, A. N. and Moore, M. K. (1977) Imitation of facial and manual gestures by human neonates. *Science*, No. 198, 75-78.

Mendoza, D. and Wichman, H. (1978) "Inner" darts: effects of mental practice on performance of dart throwing. *Perceptual and Motor Skills*, **47**, 1195-1199.

Minsky, M. (1963) Steps towards artificial intelligence. In Feigenbaum, E. A. and Feldman, J. (eds.) *Computers and Thought*, pp406-450. New York: McGraw Hill.

Minsky, M. (1967) *Computation: Finite and Infinite Machines.* Englewood Cliffs, N.J.: Prentice-Hall.

Minsky, M. (1975) Frame-system theory. In Schank, R. and Nash-Webber, B. L. (eds.) *Theoretical Issues in Natural Language Processing*, Cambridge, Mass., June 10-13, 1975.

Morris, D. (1977) *Manwatching.* London: Cape.

Moyer, R. S. (1973) Comparing objects in memory: evidence suggesting an internal psychophysics. *Perception and Psychophysics*, **13**, 180-184.

Newell, A. and Simon, H. (1972) *Human Problem Solving.* Englewood Cliffs, N.J.: Prentice-Hall.

Newell, A. (1981) Reasoning, Problem Solving and Decision Processes: The Problem Space as a Fundamental Category. In Nickerson, R. (ed.) *Attention and Performance*, Volume 8. Hillsdale, New Jersey: Erlbaum.

Nunberg, G. (1978) *The Pragmatics of Reference.* Bloomington, Indiana: Indiana University Linguistics Club.

O'Shea, T. and Young, R. M. A production rule account of errors in children's subtraction. Working Paper No. 42, Department of Artificial Intelligence, University of Edinburgh, Edinburgh, Scotland.

Paivio, A. (1971) *Imagery and Verbal Processes.* New York: Holt, Rinehart and Winston.

Paivio, A. (1975) Perceptual comparisons through the mind's eye. *Memory and Cognition*, **3**, 635-648.

Pezzoli, J. A. and Frase, L. T. (1968) Mediated facilitation of syllogistic reasoning. *Journal of Experimental Psychology*, **78**, 228-232.

Piaget, J. (1937) *The Construction of Reality in the Child.* London: Routledge and Kegan Paul. (English publication: 1955).

Pinker, S. (1984) Visual cognition. *Cognition*, **18**, 1-63.

Potter, M. (1975) Meaning in visual search. *Science*, **187**, 965-966.

Pylyshyn, Z. W. (1973) What the Mind's Eye Tells the Mind's Brain: A Critique of Mental Imagery. *Psychological Bulletin*, **80**, 1-24.

Pylyshyn, Z. (1979) The rate of "mental rotation": a test of a holistic analogue hypothesis. *Memory and Cognition*, **86**, 383-394.

Pylyshyn, Z. (1979) Validating computational models: a critique of Anderson's indeterminacy of representation claim. *Psychological Review*, **86**, 383-394.

Pylyshyn, Z. W. (1980) Computation and cognition: issues in the foundations of cognitive science. *Behavioural and Brain Sciences*, **3**, 111-132.

Pylyshyn, Z. (1981) The imagery debate: Analogue media versus tacit knowledge. *Psychological Review*, No. 88, 16-45.

Reichgelt, H. (1986) Reference and Quantification in the Cognitive View of Language. PhD Thesis, School of Epistemics, University of Edinburgh.

Revlis, R. (1975) Two models of syllogistic reasoning: fature selection and conversion. *Journal of Verbal Learning and Verbal Behaviour*, **14**, 180-195.

Roberge, J. A. (1970) A reexamination of the interpretation of errors in formal syllogistic reasoning. *Psychonomic Science*, **19**, 331-333.

Rosch, E. (1975) Family resemblance: studies in the internal structure of categories. *Cognitive Psychology*, **7**, 573-605.

Rosch, E. (1977) Classification of real-world objects: origins and representations in cognition. Chapter 13 in Johnson-Laird, P. N. and Wason, P. C. (eds.) *Thinking: Readings in Cognitive Science*, pp212-222. Cambridge: Cambridge University Press.

Russell, B. (1946) *A History of Western Philosophy*. London: Unwin.

Ryan, E. D. and Simons, J. (1982) Efficacy of mental imagery in enhancing mental rehearsal of motor skills. *Journal of Sports Psychology*, **4**, 41-51.

Samuel, A. (1981) Phoneme restoration: insights from a new methodology. *Journal of Experimental Psychology: General*, **110**, 474-494.

Sanford, A. J. and Garrod, S. C. (1981) *Understanding Written Language*. Chichester: John Wiley and Sons.

Schank, R. C. and Abelson, R. P. (1977) Scripts, Plans, and Knowledge. In Johnson-Laird, P. N. and Wason, P. C. (eds.) *Thinking*. Cambridge: Cambridge University Press.

Scribner, S. (1977) Modes of thinking and ways of speaking: culture and logic considered. In Johnson-Laird, P. N. and Wason, P. C. (eds.) *Thinking*. Cambridge: Cambridge University Press.

Sells, S. B. (1936) The atmosphere effect: an experimental study of reasoning. *Archives of Psychology*, **29**, 3-72.

Shepard, R. N. and Metzler, J. (1971) Mental rotation of three-dimensional objects. *Science*, **171**, 701-703.

Shepard, R. N. (1981) Psychophysical complementarity. In Kubovy, M. and Pomerantz, J. (eds.) *Perceptual Organisation*. Hillsdale, N.J.: Erlbaum.

Shortliffe, E. H. and Buchanan, B. G. (1976) A model of inexact reasoning in medicine. *Mathematical Biosciences*, **23**, 351-379.

Simpson, M. E. and Johnson, D. M. (1966) Atmosphere and conversion errors in syllogistic reasoning. *Journal of Experimental Psychology*, **72**, 197-200.

Sloman, A. and Hardy, S. (1983) Poplog: A Multi-Purpose Multi-Language Program Development Environment. *AISB Quarterly*, **47**, 26-34.

Smoke, K. L. (1932) An objective study of concept formation. *Psychological Monographs*, **42**.

Stamm, J. S. (1969) Electrical stimulation of monkey's prefontal cortex during delayed response performance. *Journal of Comparative and Physiological Psychology*.

Steedman, M. J. and Johnson-Laird, P. N. (1980) The Production of Sentences, Utterances and Speech Acts: Have Computers Anything to Say?. Chapter 5 in Butterworth, B. (ed.) *Language Production*, Volume 1: *Speech and Talk*. London: Academic Press.

Stenning, K. (1977) On remembering how to get there: how we might want something like a map. In Lesgold, A. M., Pellegrino, J. W., Fokkema, S. W. and Glaser, R. (eds.) *Cognitive Psychology and Instruction*, pp101-110. New York: Plenum Press.

Swinney, D. A. (1979) Lexical access during sentence comprehension: (re)consideration of context effects. *Journal of Verbal Learning and Verbal Behaviour*, **18**, 645-660.

Warren, R. (1970) Perceptual restoration of missing speech sounds. *Science*, **167**, 392-393.

Wason, P. C. (1965) The contexts of plausible denial. *Journal of Verbal Learning and Verbal Behaviour*, **4**, 7-11.

Wason, P. C. (1966) Reasoning. In Foss, B. (ed.) *New Horizons in Psychology*. Harmondsworth, Middlesex: Penguin.

Wason, P. C. (1972) In real life, negatives are false. *Logique and Analyse*, **19**, 19-38.

Wason, P. C. (1977) Self-contradicions. Chapter 7 in Johnson-Laird, P. N. and Wason, P. C. (eds.) *Thinking: Readings in Cognitive Science*, pp114-128. Cambridge: Cambridge University Press.

Wason, P. C. and Evans, J. S. B. T. (1975) Dual processes in reasoning?. *Cognition*, **3**, 141-154.

Wilkins, M. C. (1928) The effect of changed material on the ability to do formal syllogistic reasoning. *Archives of Psychology*, **16**.

Winograd, T. (1972) *Understanding Natural Language*. New York: Academic Press.

Wittgenstein, L. (1953) *Philosophical Investigations*. Oxford: Basil Blackwell. Translated by G. E. M. Anscombe.

Woodworth, R. S. and Sells, S. B. (1935) An atmosphere effect in formal syllogistic reasoning. *Journal of Experimental Psychology*, **18**, 451-460.

Young, R. M. (1973) Children's Seriation Behaviour: A Production System Analysis.. PhD Thesis, Carnegie Mellon University.

Zloof, M. M. (1977) Query-by-example: a database language. *IBM Systems Journal*, 324-343.

Zurif, E. and Blumstein, S. (1978) Language and the brain. Chapter 6 in Halle, M., Bresnan, J. and Miller, G. (eds.) *Linguistic Theory and Psychological Reality*, pp229-246. Cambridge, Mass.: MIT Press.

APPENDIX A

Results from the Author's Experiment

Explanation of Symbols

All tables assume a syllogism relating a and b in the first premise and b and c in the second.

Premises and conclusions are encoded according to the following scheme:

"A(x, y)" encodes "all x are y".
"E(x, y)" encodes "no x are y".
"I(x, y)" encodes "some x are y".
"O(x, y)" encodes "some x are not y".
"nvc" encodes "no valid conclusion"
"with y" encodes a response involving the middle term
"trivial" encodes a "deduction" that follows from only one premise
– e.g. **O(q, r)** in response to a premise pair containing **E(q, r)**.
"other" encodes a response that could not be interpreted as any syllogistic form

The figures at the foot of each entry indicate the mean response time for subject on that premise pair and their average confidence in their answer.

Figure 1

2nd Premise	1st Premise A(x, y)		1st Premise I(x, y)		1st Premise E(x, y)		1st Premise O(x, y)	
A(y, z)	A(x, z)	13	I(x, z)	15	nvc	6	nvc	1
	A(z, x)	2	I(z, x)	2	E(x, z)	13	I(x, z)	7
	I(x, z)	5	O(x, z)	1	E(z, x)	1	I(z, x)	3
	with y	1	O(z, x)	1	O(z, x)	1	E(z, x)	1
			with y	2			O(x, z)	8
							O(z, x)	1
	35 Sec	Conf=7	33 Sec	Conf=7	45 Sec	Conf=7	45 Sec	Conf=7
I(y, z)	nvc	3	nvc	12	nvc	8	nvc	15
	I(x, z)	16	I(x, z)	7	E(x, z)	5	I(x, z)	5
	with y	2	I(z, x)	1	E(z, x)	1	O(z, x)	1
			O(x, z)	1	O(x, z)	1		
					O(z, x)	5		
					with y	1		
	30 Sec	Conf=6	40 Sec	Conf=6	44 Sec	Conf=6	54 Sec	Conf=5
E(y, z)	nvc	1	nvc	7	nvc	17	nvc	9
	E(x, z)	17	I(x, z)	1	E(x, z)	3	I(x, z)	1
	E(z, x)	2	E(x, z)	4	E(z, x)	1	I(z, x)	1
	O(x, z)	1	O(x, z)	8			E(x, z)	3
			with y	1			O(x, z)	6
							O(z, x)	1
	28 Sec	Conf=7	44 Sec	Conf=6	32 Sec	Conf=6	64 Sec	Conf=5
O(y, z)	nvc	5	nvc	11	nvc	16	nvc	15
	I(x, z)	2	I(x, z)	2	A(x, z)	1	I(x, z)	2
	O(x, z)	14	O(x, z)	6	E(x, z)	2	O(x, z)	2
			O(z, x)	2	E(z, x)	1	with y	2
					O(z, x)	1		
	32 Sec	Conf=7	44 Sec	Conf=5	51 Sec	Conf=5	40 Sec	Conf=5

Figure 2

2nd Premise	1st Premise A(y, x)		1st Premise I(y, x)		1st Premise E(y, x)		1st Premise O(y, x)	
A(z, y)	nvc	2	nvc	5	nvc	1	nvc	4
	A(x, z)	1	I(x, z)	2	E(x, z)	1	I(x, z)	1
	A(z, x)	15	I(z, x)	14	E(z, x)	16	I(z, x)	2
	I(z, x)	2			O(z, x)	1	E(z, x)	1
	with y	1			with y	2	O(z, x)	12
							with y	1
	30 Sec	Conf=8	33 Sec	Conf=6	41 Sec	Conf=7	32 Sec	Conf=6
I(z, y)	nvc	1	nvc	15	nvc	4	nvc	12
	I(x, z)	2	I(z, x)	5	I(z, x)	1	I(z, x)	1
	I(z, x)	16	with y	1	E(x, z)	1	O(z, x)	4
	with y	2			E(z, x)	3	with y	4
					O(z, x)	11		
					other	1		
	32 Sec	Conf=7	31 Sec	Conf=7	31 Sec	Conf=6	36 Sec	Conf=6
E(z, y)	nvc	7	nvc	10	nvc	16	nvc	13
	E(x, z)	3	E(x, z)	1	A(z, x)	1	E(x, z)	1
	E(z, x)	8	E(z, x)	5	I(z, x)	1	E(z, x)	6
	O(x, z)	1	O(x, z)	3	E(z, x)	3	O(x, z)	1
	with y	2	O(z, x)	1				
			with y	1				
	38 Sec	Conf=7	45 Sec	Conf=6	53 Sec	Conf=5	48 Sec	Conf=5
O(z, y)	nvc	3	nvc	11	nvc	10	nvc	13
	I(x, z)	2	I(z, x)	2	I(x, z)	2	I(x, z)	1
	I(z, x)	10	O(x, z)	2	I(z, x)	2	I(z, x)	2
	O(x, z)	1	O(z, x)	3	E(x, z)	1	O(z, x)	4
	O(z, x)	4	with y	3	E(z, x)	2	with y	1
	other	1			O(z, x)	4		
	50 Sec	Conf=6	58 Sec	Conf=5	44 Sec	Conf=6	59 Sec	Conf=5

Figure 3

2nd Premise	1st Premise A(x, y)		I(x, y)		E(x, y)		O(x, y)	
A(z, y)	nvc	11	nvc	8	nvc	4	nvc	8
	A(x, z)	4	I(x, z)	7	E(x, z)	9	I(x, z)	3
	A(z, x)	2	I(z, x)	5	E(z, x)	6	I(z, x)	3
	I(z, x)	1	O(x, z)	1	with y	2	O(x, z)	3
	with y	3					O(z, x)	3
							with y	1
	35 Sec	Conf=7	46 Sec	Conf=6	39 Sec	Conf=7	38 Sec	Conf=6
I(z, y)	nvc	8	nvc	15	nvc	6	nvc	13
	I(x, z)	2	I(x, z)	3	A(x, z)	1	I(x, z)	3
	I(z, x)	10	I(z, x)	2	I(z, x)	1	O(x, z)	3
	with y	1	O(x, z)	1	E(x, z)	3	with y	2
					O(x, z)	2		
					O(z, x)	7		
					with y	1		
	38 Sec	Conf=6	25 Sec	Conf=6	54 Sec	Conf=6	53 Sec	Conf=5
E(z, y)	nvc	4	nvc	6	nvc	18	nvc	12
	E(x, z)	7	I(x, z)	2	E(x, z)	1	I(x, z)	3
	E(z, x)	8	E(x, z)	3	E(z, x)	2	I(z, x)	2
	O(x, z)	1	E(z, x)	4			E(x, z)	1
	with y	1	O(x, z)	3			E(z, x)	1
			with y	3			O(x, z)	2
	35 Sec	Conf=6	40 Sec	Conf=6	38 Sec	Conf=6	47 Sec	Conf=6
O(z, y)	nvc	4	nvc	14	nvc	11	nvc	14
	I(x, z)	3	I(x, z)	2	I(z, x)	3	I(x, z)	3
	I(z, x)	2	I(z, x)	1	E(z, x)	2	I(z, x)	1
	O(x, z)	3	O(x, z)	2	O(z, x)	3	O(x, z)	2
	O(z, x)	5	O(z, x)	1	with y	2	with y	1
	with y	4	with y	1				
	50 Sec	Conf=6	38 Sec	Conf=6	66 Sec	Conf=5	42 Sec	Conf=6

Figure 4

2nd Premise	1st Premise A(y, x)		I(y, x)		E(y, x)		O(y, x)	
A(y, z)	nvc	5	nvc	1	nvc	6	nvc	2
	A(x, z)	8	I(x, z)	7	E(x, z)	4	A(x, z)	2
	I(x, z)	5	I(z, x)	12	E(z, x)	7	I(x, z)	1
	I(z, x)	3	with y	1	O(z, x)	3	I(z, x)	3
					trivial	1	O(x, z)	1
							O(z, x)	12
	33 Sec	Conf=7	20 Sec	Conf=7	44 Sec	Conf=7	45 Sec	Conf=7
I(y, z)	nvc	1	nvc	13	nvc	9	nvc	12
	I(x, z)	11	I(x, z)	3	E(x, z)	2	I(x, z)	3
	I(z, x)	7	I(z, x)	3	E(z, x)	4	O(z, x)	6
	with y	2	E(x, z)	1	O(z, x)	6		
			with y	1				
	28 Sec	Conf=7	40 Sec	Conf=6	32 Sec	Conf=6	38 Sec	Conf=6
E(y, z)	nvc	6	nvc	9	nvc	20	nvc	10
	E(x, z)	6	E(x, z)	2	A(x, z)	1	A(z, x)	1
	E(z, x)	4	E(z, x)	1			E(x, z)	2
	O(x, z)	3	O(x, z)	8			E(z, x)	3
	O(z, x)	1	O(z, x)	1			O(x, z)	2
	with y	1					trivial	1
							with y	2
	32 Sec	Conf=6	51 Sec	Conf=6	56 Sec	Conf=5	55 Sec	Conf=5
O(y, z)	nvc	2	nvc	13	nvc	14	nvc	14
	I(x, z)	2	I(x, z)	1	I(x, z)	2	I(x, z)	2
	O(x, z)	15	I(z, x)	1	E(z, x)	1	O(x, z)	4
	O(z, x)	1	E(x, z)	1	O(z, x)	3	O(z, x)	1
	with y	1	O(x, z)	5	with y	1		
	43 Sec	Conf=7	26 Sec	Conf=6	75 Sec	Conf=5	34 Sec	Conf=6

APPENDIX B

Paper Presented at AISB Conference, 1985.

The table refered to as Table 1 in this paper appears as Table 4.1, P106.
The table refered to as Table 2 in this paper appears as Table 7.1, P234.
The table refered to as Table 3 in this paper appears as Table 7.2, P236.

Modelling Syllogistic Reasoning Using Simple Mental Models.

Robert Inder

Department of Artificial Intelligence and School of Epistemics,
University of Edinburgh.

Syllogistic reasoning is a rich field within psychological research, and is one of the key topics in the formulation of Johnson-Laird's theories of Mental Models. The nature of the task is outlined, and a brief summary of the main theories in the area is presented, including Johnson-Lairds own account. An alternative theory is presented which is based on using surface features of the premises in conjunction with a simplified form of mental model. Individual subjects are modelled as possessing specific groups of simple "strategies", and a computer model is used to demonstrate the adequacy of the approach.

INTRODUCTION.

One term that is currently attracting a great deal of attention among those studying cognition is "mental model", although there is a broad spectrum of meanings that seem to be associated with it. At one extreme it is used merely to refer to a block of knowledge that the subject appears to possess, with no stronger claim explicitly made, a style of usage summarised by the suggestion that a conceptual model is "a more or less definite representation or metaphor that a user adopts to guide his actions and help him interpret a device's behaviour" (Young 1983, P35). One of the best known workers at the other end of the spectrum is Johnson-Laird, who, apart from having written a book on the subject (Johnson-Laird 1983), has been supporting the concept for many years (Johnson-Laird 1975, 1980). He suggests that mental models play a fundamental role (conscious or otherwise) in the processes by which people understand, remember and think about the world, and uses them to argue against the idea a "logic in the head". However, Johnson-Laird claims that there are two types of mental models: "Physical models represent the physical world; conceptual models represent more abstract matters" (Johnson-Laird 1983, P422). Physical mental models have direct experimental support (e.g. Mani, K. and Johnson-Laird, P. N., 1982) and strong connection with both experiments and intuitions relating to mental images, while to justify postulating conceptual models, Johnson-Laird calls upon evidence from tasks in syllogistic reasoning.

This paper will present, in the context of Johnson-Laird's own theory, a new approach to the explanation of syllogistic reasoning and describe the use of a computer program to demonstrate its accuracy. While inspired by an early version of Johnson-Laird's theory (Johnson-Laird and Steedman, 1978), the new theory suggests a way in which the experimental data can be explained with recourse only to physical mental models. Finally, it should be borne in mind that the objective is to understand how untrained people go about solving syllogistic problems, and NOT to say anything new about their logical properties or their solution by computer.

WHAT ARE SYLLOGISMS.

Consider a statement of the form "Some artists are beekeepers". One of the ways it can be interpreted is as expressing a relation between sets of individuals : in this case, that within a certain domain, the intersection of the set of artists with the set of beekeepers is not empty. Similar readings follow naturally for related sentences obtained by replacing "Some" by "All" or "None", or "are" by "are not". This gives four possible sentence "shapes":

"All of the whatsits are doobries"	(known as A, from Affirmo)
"Some of the whatsits are doobries"	(known as I, from affIrmo)
"None of the whatsits are doobries"	(known as E, from nEgo)
"Some of the whatsits are not doobries"	(known as O, from negO)

If two such statements are considered then, provided they have a predicate in common, it may be possible to combine them to deduce a third sentence of a similar form. Thus, given the premises "All artists are beekeepers" and "All beekeepers are chemists", one can deduce that "All artists are chemists". Combination of such premises is the basis of the categorial syllogism, with two premises, each in one of four forms, giving a range of 16 possible combinations.

In addition to the forms of the individual premises, there is another degree of freedom within a syllogism, namely the order of the predicates within the premises. Thus for the case of a syllogism relating As, Bs and Cs, there are four possible arrangements of the premises:

1.	2.	3.	4.
A - B	B - A	A - B	B - A
B - C	C - B	C - B	B - C

Following Johnson-Laird, this will be referred to as the "figure" of the syllogism, with the numbering shown above serving to name the figures. However, this usage deserves to be hedged. In traditional terminology, the syllogism is associated with a specific order of terms in its conclusion, and the identification of the figure from the arrangement of the terms within the premises takes account of the roles that they will play in the conclusion. This gives a similar but different numbering of possible term arrangements. Taking both these factors into account, the fact that there are 64 possible syllogistic problems, 16 in each of four figures, should be reasonably obvious. These can be referred to by indicating the figure and letters corresponding to the forms of the premises: thus "2EI" refers to the syllogism in the second figure, with the first premise being in "E" form, and the second being "I" -- i.e. No Bs are As, Some Cs are Bs. Of the 64 possible premise combination, 27 have valid conclusions, which are indicated in Table 1.

Having looked at the "right" answers, the next step is to consider the kind of answers people actually give. Johnson-Laird has reported two experiments where subjects have been asked to give conclusions to every possible syllogism (Johnson-Laird and Steedman, 1978, and Johnson-Laird and Bara, 1984, also presented in Johnson-Laird 1983), and recent work at Edinburgh (Inder, forthcoming) has given basically similar results. The first thing to point out about the results of these experiments is that they show that untrained people are not very good at syllogisms, with typical individual scores between 50% and 75%. However, the most striking feature is that although there are nine possible conclusions that subjects could offer to each pair of premises (four forms in each of two orders, and no valid conclusion), the actual distribution of responses is far from even: most problems have a clearly favoured response (or pair of responses), which may or may not be valid.

In addition, one of the most repeatable features of all the experimental results is the figural effect, discovered by Johnson-Laird. Whenever a conclusion of the form "some (or no) X are Y" is valid, so also is a conclusion of the form "some (or no) Y are X": they are logically identical. However, subjects show a strong preference between the two forms depending on the figure of the premises. With premises in the first figure (i.e. A - B, B - C) subjects prefer to offer conclusions in the form A - C, while in the fourth figure (B - A, C - B) they prefer C - A. In the remaining ("symmetrical") figures, subjects show no preference. This effect appears both with valid and invalid conclusions, and in individual subjects. In other words, even where subjects are offering incorrect (invalid) conclusions, they are not doing so at random but are clearly working to some pattern. The problem is to determine the nature of this pattern.

THEORIES OF SYLLOGISTIC REASONING

For present purposes it is sufficient only to touch on the (extensive) literature on the psychology of syllogistic reasoning (For a more thorough review, and copious references, see (Evans, 1982)). As far as explaining the mechanisms underlying syllogistic reasoning, there are two main theories: the atmosphere effect, and illicit conversion. Unsurprisingly, since they are the dominant theories in the field, both approaches are supported by reasonable experimental results, usually with minor but possibly significant methodological variations.

The atmosphere effect was first proposed by Woodworth and Sells (1935), working with data from validation experiments (Sells, 1936, and Wilkins, 1928). They

suggested that the quantifiers employed in the premises tend to create an "atmosphere", and that subjects are more inclined to support conclusions with quantifiers that are compatible with that atmosphere. For instance, a syllogism consisting of solely universal premises (All (A) and no (E)) will have a "universal" atmosphere, and subjects will be inclined to accept a universal conclusion. Similarly if the premises were both positive (All or some) subjects would respond to the positive atmosphere and prefer positive conclusions. Where the premises are mixed, particular and negative conclusions are preferred.

Notice that, according to the atmosphere theory, solving syllogisms is (partly) a non-rational process. Evans suggests that the unacceptability of this was a significant factor in motivating Chapman and Chapman (1959) to propose the system known as "illicit conversion". According to this, subjects are essentially rational, in that they attempt to find a valid conclusion using sound logical methods. However, they are led astray by the fact that they mis-represent the premises. Thus when dealing with a premise such as "all artists are beekeepers", they will treat this as though it also implied that "all beekeepers are artists" -- in other words, as a statement of set equality. Similarly, premises of the form "some artists are not beekeepers" will be taken as also saying that "some beekeepers are not artists". This rational/irrational distinction is blurred somewhat by the Chapmans' introduction of "probabilistic reasoning", although despite this the distinction between these two accounts has provided a major focus for work in syllogistic reasoning.

JOHNSON-LAIRD'S THEORY OF SYLLOGISTIC REASONING

In contrast to these two approaches, Johnson-Laird offers an account based on mental models which has undergone a significant metamorphosis between its original presentation (Johnson-Laird and Steedman, 1978) and its "current" form (Johnson-Laird, 1984, and Johnson-Laird and Bara, 1984 and outlined here).

The theory is introduced by first describing how one could go about solving a syllogistic problem using a room full of actors, an idea inspired by a subject's introspections. Suppose that the premises were

"All artists are beekeepers"

and "Some beekeepers are chemists".

One could begin by asking some of the actors to take on the roles of artists in some manner, such as clutching paintbrushes and a pallette. Then one would ask them to also take on the role of beekeeper, equipping them with suitable headgear or whatever. In the light of the fact that there might also be some beekeepers who are not artists, one could also similarly equip one or two of the remaining, "unartistic", actors. At this point, the actors in some sense "represent" the first premise, and in using a notation introduced by Johnson-Laird, the situation is shown in fig. 1(i), where the brackets round the solitary "b" indicate the uncertainty whether such an individual exists.

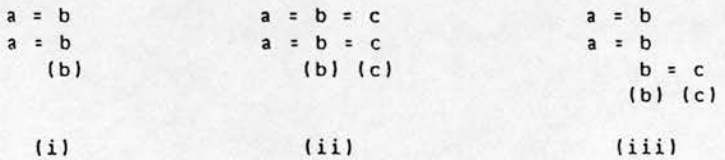


Figure 1.

The second premise could now be dealt with by issuing some of the beekeepers, maybe those with paint brushes, with test tubes, thus allowing them to adopt the role of chemist as well. This would be reflected by a diagram such as fig 1(ii). Anyone studying this situation might be inclined to suggest that "all the artists are chemists", which is true in the current situation, is a valid inference. However, this conclusion can be tested by getting the actors playing chemists to pass the test-tubes to other beekeepers, possibly to those who are not artists. Doing this we can get a situation which corresponds to fig 1(iii), from which we can clearly see that this potential conclusion need not be the case, and in fact the premises remain true even if none of the artists are also chemists. Since all, some or none of the artists may also be chemists,

there is no relation that must hold and hence no valid conclusion that can be drawn.

In the light of this intuitive description, Johnson-Laird's actual theory is very straightforward. When confronted with a pair of syllogistic premises, people construct the appropriate mental model, which Johnson-Laird represents using the notation used above. From this model, they attempt to "read off" a conclusion. If they cannot, they know that the problem has no valid conclusion. Otherwise, they may embark on a sequence of model manipulations which attempt to disprove the possible conclusion. If this succeeds, the conclusion is rejected and the process is repeated using the modified model. Otherwise, if the model cannot be altered to falsify the conclusion but not the premises, the conclusion is accepted as valid.

If this modification process is repeated correctly, subjects will produce correct answers for rational reasons (i.e. they will have found a conclusion that they know cannot be defeated), but without employing deductive reasoning: I.e. subjects have the potential for true rationality. However, experiments show that subjects are far from perfectly logical. Johnson-Laird explains their aberrations by saying that subjects often do not take the read/falsify process to completion: Instead, they break off prematurely, and are thus led to accept a conclusion that could be falsified. In the example above this would lead to subjects being inclined to accept the conclusion suggested by the original model (namely, "some artists are chemists"), and results shows that this is in fact the most common response. One of the strong points of the theory is that the conclusions that can be read off the "intermediate" models correlate well with the observed responses of experimental subjects.

Given this outline, one can appreciate a few of the details of the theory. The conclusion that subjects will read off any particular model is basically the strongest possible conclusion that is definitely true in the model. Negation is represented by the presence of a "barrier" in the model. Thus figure 2(i) shows the model that would be built in response to the 1EE syllogism, i.e.

None of the artists are beekeepers
None of the beekeepers are chemists

This could suggest "No artists are chemists", which is not valid, but is still a frequently-drawn conclusion. However, Johnson-Laird provides precise rules for model manipulation covering the movement of terms past barriers, which allow this model to be altered to give the model in figure 2(ii). Explicit negative information, as represented by the barrier, is distinct from uncertainty, as shown by the absence of a positive connection. Thus this model does not allow one to say that any artists are chemists: i.e. no conclusion can be read from it.

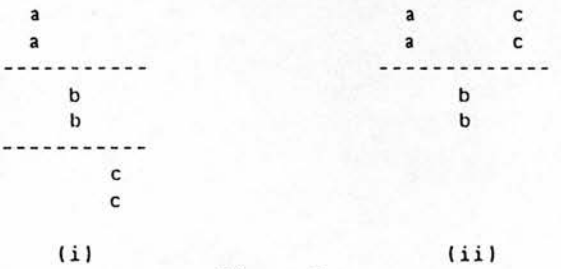


Figure 2

Johnson-Laird suggests that the figural effect arises from the way that the models are built. His theory assumes that combined mental models can only be readily built when the premises are in the first figure. In other figures, additional operations must be carried out to manipulate the models into an appropriate arrangement. He then calls upon the idea that memory is in some sense "first in, first out" to suggest that subjects will tend to offer conclusions that use terms from the model in the same order they were used during its construction. Notice that this means that the order of the properties within the model matters. There is a significant difference between these two models:

a = b	b = a
a = b	b = a
(b)	(b)

A NEW APPROACH.

It should be noticed that all the theories mentioned so far have been formulated to describe the patterns of responses observed within a population of subjects. Johnson-Laird has commented on the differences between individuals. Variability of response, in his theory, is caused by the range of possible points at which subjects break out of the conclusion testing loop. For the most part he suggests that this is caused by limitations on working memory, although he does suggest that other factors are relevant. In contrast, the theory to be presented here is centred around the detailed study of individual response patterns, involving computational models to demonstrate the descriptive adequacy of the internal mental operations it postulates.

According to the new theory, non-logicians solve syllogistic problems by the interaction of two mechanisms: one which generates potential solutions, and one which attempts to falsify them. As with Johnson-Laird's theory, it is assumed that reading the premises results in a mental model, although there are three significant differences. Firstly, his models employ "optional" (bracketed) entities to attempt to cover the range of possible situations, which have no counterpart in the new theory. Secondly, his theory depends on the fact that his models distinguish between ignorance and definite negative knowledge, which is explicitly marked. The models used in the new theory do not make this distinction: an entity in the model either has or does not have any particular property. These differences arise from the fact that Johnson-Laird's models in some sense capture the truth conditions of the premises, and need the optional entities and explicit negation to do this. In contrast, the new theory suggests that the models that are built represent a single, definite situation which for the model builder is typical of the kind of situation that the premises are describing. Finally, his theory imposes constraints on the "syntax" of the storage of the models which are not necessary in the new approach.

Once the premises are modelled, the first of the two solving processes operates on (something very close to) the surface form of the premises, and produces a sequence of words that is a possible conclusion. Most people unfamiliar with syllogistic reasoning will not generate candidate solutions methodically, but will typically have one or two strategies for producing a conclusion, possibly as a result of their (very limited) encounters with quantified sentences in everyday life. If no conclusion-suggesting process can produce anything from the premises in their original form, they will be manipulated, typically by a simple shuffling of the words and without any particular regard for logic or truth. Once a candidate solution is produced, the model created by the premises is checked to see whether the "conclusion" is true. If it is not, then clearly it is not a valid conclusion (the model represents a counter-example) and is rejected. If a solution holds in the initial model, the second of the solving processes may come into effect. This will result in an attempt to modify the model in order to defeat the potential conclusion. As with conclusion generation, this is not a normal activity, so subject's procedures will be disorganised and inefficient. Indeed, some subjects can be modelled as performing no testing at all. However, if testing occurs and succeeds, the conclusion is once again rejected.

If a potential conclusion cannot be disproved by any of the model manipulation techniques that the subject can employ, it will be accepted as valid. On the other hand, if it is rejected, the first procedure will be called upon to produce another. If it cannot, the problem will be treated as having no conclusion. If it can, the cycle of testing begins again. With experience, subjects will develop better (basically more thorough and methodical) strategies both for generating and defeating potential conclusions. The end point of this process would be a subject able to quickly construct an appropriate counter example for any possible conclusion, and to generate the right conclusion for any specific pair of premises, making validation unnecessary (i.e. they would know the answers off by heart). In particular, subjects would learn to directly recognise (rightly or wrongly) that certain syllogisms do not have a conclusion, in which case the first

procedure would generate "no conclusion" directly.

COMPUTER MODELLING

The essential feature of this theory is the claim that the complex patterns observed in subjects responses to syllogistic problems can be explained by the interaction of a few simple rules, which are in themselves essentially "plausible". It has a clear family resemblance to the work carried out by Brown and Burton (1978) on the analysis of children's subtraction skills in terms of interacting sub-skills. It also has a certain similarity with the basic activity of linguistics, where building a grammar is very much a matter of finding sets of simple elements (transformations and filters, or categories for words) which combine to produce the observed range of language behaviour. In work of this kind, it is not always easy to tell from the rules alone whether a particular specimen can be explained. For this reason, a computer is used to "animate" the sub-skills, grammar rules or solution strategies, and thus clarify the extent of the fit with the actual observations. In the case of the new theory, the program consists of about 1000 lines of (distinctly under-commented) prolog code, with a further 200 lines or so potentially specific to capturing each individual subject. It takes between 6 and 10 seconds of CPU time on a VAX 11/750 to process each syllogism, which translates to between 10 seconds and a minute per syllogism, or between fifteen minutes and an hour and a half to produce a conclusion for an entire batch of sixty-four.

Within the simulator there are, obviously, many routines that are of no interest: they perform such tasks as controlling the program through a batch of syllogisms, reading premises and writing conclusions. The highest routine that has any bearing on the present theory is the top-level solving routine, which is virtually trivial:

```
solveit(Premises,Conclusion) :-
    generate_conclusion(Premises,Conclusion),
    not falsify_conclusion(Premises,Conclusion),
    !.
solveit(Premises,nvc).
```

Roughly, this can be glossed as "the subject suggests Conclusion if they can generate it from the Premises, and not falsify it. Otherwise, they will say there is no valid conclusion". The prolog mechanisms for backtracking (see Clocksin and Mellish, 1981) ensures that, should a generated conclusion be falsified, "generate_conclusion" will be re-entered to attempt to generate another. Clearly this level says very little of interest and it is necessary to consider the next level of detail, namely, the activities within "generate_conclusion" and "falsify_conclusion".

The basic function of "generate_conclusion" is to present (rearrangements of) the premises to "suggest_conclusion", and to check that the conclusion offered is "acceptable" (N.B. not valid) in the sense of being in a syllogistic form and not containing the middle term. It is "suggest_conclusion" that actually offers potential conclusions in response to the premises in their current arrangement, using such strategies as:

```
suggest_conclusion([Premise1,[all, Bs, are, Cs]], Conclusion)
        :-  replace(Bs, Cs, Premise1, Conclusion).
```

This can be glossed as "if the second premise has the form 'All, Bs, are, Cs', replace any occurrence of 'B' in the first premise by 'C'" and seems about as obvious a strategy as one can get.

To complement "generate_conclusion", "falsify_conclusion" passes the putative conclusion (and the premises, although this has not actually been used yet) to "modify_strategy", which attempts to modify the model to defeat the conclusion, and then checks to see whether in fact a counter example has been created. Typically, "modify_strategy" will employ such rules as:

```
modify_strategy(Premises,[no,A,are,B]) :- true([some,A,are,B],Make).
```

This can be glossed as "when trying to defeat conclusion of the form 'no As are Bs', try to make 'some As are Bs' true".

AN EXAMPLE SUBJECT

Having described the theory and the machinery, there is enough context in place to consider some results. Table 2 contains the results given by one particular subject (Subject A). One response (2EI) is considered as an "aberration" since the subject's actual response could not be unambiguously categorised into one of the requested forms: Henceforth, it is treated as a response of "nvc" ("no valid conclusion"). The conclusions given are shown in compressed form in table 3(i).

The first feature to notice about the responses is that the 28 premise pairs containing an "A" premise provide 20 of the 23 conclusions offered, including all of the invalid ones. This suggests that an "A" premise might well facilitate conclusion offering, an idea that was first considered by a workshop in the Edinburgh University School of Epistemics and discussed more fully under the heading of "the A-effect" in Lee (1983). In fact, the example conclusion generation rule offered above (substituting predicates from an "all" premise into another "congruent" premise), when combined with very straightforward model making and some means of rearranging the premises to try all possible orderings, gives 22 of the 28 responses correctly, and another three correct except for the predicate order (see table 3(ii)). Adding enough skill to try "some X are Y" when "all X are Y" has been found inappropriate corrects 4AA, and leaves three types of response to be explained: conclusions offered for problems not involving "A" premises, the "E" conclusions at 4AE and 2AE and the "nvc" at 30A.

When seeking a strategy for getting conclusions from premise pairs that do not have an "A" premise, there is a significant problem. All the conclusions offered are for premise pairs containing an "I" and an "E". However, although all eight such premise pairs permit a conclusion, the subject only offers one for three of them. It would undoubtedly be possible to find some combination of rules that generate all and only these three solutions. However, the subject's solution-seeking arsenal currently includes the strategy of considering all possible permutations of premises and predicates. If nothing more is done, the subject would be expected to rearrange the remaining five premise pairs until they matched the rules, and thus produce a valid (undefeatable) conclusion. Thus the model of the subject would have to be revised to limit the ability to re-order the premises, and the "A-effect" rule made correspondingly more complicated in order to compensate.

However, there is another feature of the subject's responses to the eight combinations of E and I premises: as the subject dealt with these eight (one at a time, scattered at random throughout the total batch), a conclusion was only offered for the LAST THREE. This strongly suggests that, rather than possessing a specific, asymmetrical strategy that works only for some such combinations of premises, the subject uses a general strategy that will work for any of them, but only ACQUIRES it two thirds of the way through the experiment, as a result of the experience of attempting the first 40 or more examples. If this idea is accepted, then there is no need to revise the treatment of premise pairs containing "all" since the new rule is itself symmetric.

Consider the remaining inaccuracies of the computer model. The invalid "E" conclusions should be no surprise, since the subject as modelled so far makes no attempt (or has no idea how) to disprove tentative conclusions. Simply adding some skill at testing "E" conclusions, by trying to make a falsifying group of individuals, will easily correct them both. Unfortunately, it will also defeat the desired (but invalid) "E" conclusions at 1EA and 4EA. However, the data in Table 2 reveal that the subject spent 110 seconds studying the 2AE problem before deciding that there was no conclusion. It therefore seems reasonable to suggest that the result of this considerable effort was the recognition of the method of attempting to falsify "E" conclusions. If this assumption is made then the unwanted responses to 4AE and 2AE are prevented while 1EA remains. Unfortunately this also suggests that the subject should have been able to defeat the "E" conclusion which was in fact offered in response to the 4EA, and this discrepancy is not easy to fix and remains a misfit between the model and the subject.

The unwanted conclusion that the model offers for 3OA is valid, so it cannot be removed by sharpening up the subject's logical ability (which is also why the unwanted conclusions to combinations of I and E premises had to be dealt with by limiting the generation of candidate solutions). In fact, quite the opposite, the aim is to deliberately introduce illogicality. One method of doing this is to postulate that the "all" substitution cannot occur into predicates of the form "some A are not B", which is in effect suggesting that, initially at least, the subject took the premises to be "some A are not-B" and "all C are B", and did not utilise the connection between "not-B and "B".

If the suggestions put forward above are followed, the set of rules one arrives at produce the solutions presented in Table 3(iii). These differ from the actual responses in two ways: The subject erroneously offered a "E" conclusions to the 4EA problem while the program does not, and the model reverses the premise order in three conclusions (4IA, 3EA, 3AE).

In the light of this model, how can the subject's performance be described? Initially the only strategy available for formulating a possible conclusion was to substitute one predicate for another when prompted to do so by an "all" premise. This modest initial ability is not surprising if one considers that inferring one sentence from a pair of others in the absence of a sensible context is NOT an everyday activity. Similarly, since disproving deductions is also unfamiliar, the subject has no effective way of doing this either. After twenty problems, 153 seconds spent on the 1EO problem leads to the realisation of how to exploit the connection between B and not-B. Another twenty-problems later, and 110 seconds playing with the 2AE leads to a recognition of what to do to falsify an "E" conclusion, although it is not enough to uncover the possibility of trying an "O" conclusion (which would have been valid) instead. Finally, soon after that the subject discovers that an "E" premise and an "A" premise can combine to give an "O" conclusion. At the end of the hour, the subject has improved in this unfamiliar formal task.

CONCLUDING REMARKS.

This new approach to explaining subjects' syllogism solving represents an attempt to satisfy Johnson-Laird's goal of explaining reasoning behaviour without postulating any kind of "logic in the head". The mental models it uses inherit the intuitive appeal and experimental support of Johnson-Laird's physical models, but dispense with the necessity for specifically restricted (sequential) access and express negation that complicate his conceptual models. Further, the general structure has widespread intuitive appeal and clearly indicates how a particular subject progresses towards expertise in what is treated as essentially a problem-solving exercise. The next step is to produce models for a range of subjects, and attempt to build up a picture of the common features of individual's initial strategies and learning mechanisms, an exercise that has essentially the same flavour as the linguists' attempts to discover the underlying mechanisms of human language processing by producing grammars for a range of languages.

REFERENCES:

Brown and Burton (1978) Diagnostic Models for Procedural Bugs in Basic Mathematical Skills. Cognitive Science, 2, 155 - 192.

Chapman, I. J. and Chapman, J. P. (1959) Atmosphere effect re-examined. Journal of Experimental Psychology, 58, 220-226.

Clocksin, W. F. and Mellish, C. (1981) Programming in Prolog. Springer-Verlag.

Evans, J. St. B. T. (1982) The Psychology of Deductive Reasoning. Routledge and Kegan Paul.

Inder, R. (forthcoming) PhD. Thesis. University of Edinburgh.

Johnson-Laird, P.N. (1975) Models of Deduction. IN Falmagne, R. J. Reasoning: Representation and Process in Children and Adults. Erlbaum.

Johnson-Laird, P.N. (1980) Mental Models in Cognitive Science. Cognitive Science, 4, 71 - 115.

Johnson-Laird, P.N. (1983) Mental Models. Cambridge University Press.

Johnson-Laird, P. N. and Steedman, M. J. (1978) The Psychology of Syllogisms.

Cognitive Psychology, 10, 64 - 99.
Johnson-Laird, P. N. and Bara, B. G. (1984) Syllogistic Inference. Cognition, 16, 1-61.
Lee, J. R. (1983) Illicit Conversion and the A-Effect. Working paper, Reasoning Workshop, School of Epistemics, University of Edinburgh.
Mani, L. and Johnson-Laird, P. N. (1982) The mental representation of spatial descriptions. Memory and Cognition 10, 181 - 187.
Sells, S. B. (1936) The atmosphere effect: an experimental study of reasoning. Archives of Psychology, 29, 3-72.
Wilkins, M. C. (1928) The effect of changed material on the ability to do formal syllogistic reasoning. Archives of Psychology, 16, No. 102.
Woodworth, R. S. and Sells, S. B. An atmosphere effect in formal syllogistic reasoning. Journal of experimental Psychology, 18, 451-460
Young, R. M. Surrogates and Mappings: Two Kinds of Conceptual Models for Interactive Devices. IN Gentner, D. and Stevens, A.L. (eds) Mental Models. Erlbaum.